Estimation non paramétrique adaptative dans la théorie des valeurs extrêmes : application en environnement

Quang Khoai Pham

THESE / UNIVERSITE DE BRETAGNE-SUD
sous le sceau de l'Université européenne de Bretagne

pour obtenir le titre de
DOCTEUR DE L'UNIVERSITE DE BRETAGNE-SUD
Mention : Mathématiques
Ecole doctorale : SICMA

présentée par
Quang-Khoai Pham
Laboratoire de Mathématiques de Bretagne
Atlantique UMR CNRS 6205

Estimation non paramétrique adaptative dans la théorie des valeurs extrêmes : application en environnement

Thèse soutenue le 09 janvier 2015
devant le jury composé de :

Jean-Marc Azaïs
Professeur, Université Paul Sabatier, Toulouse / Rapporteur
Jean-Noël Bacro
Professeur, Université Montpellier 2 / Examinateur
Patrice Bertail
Professeur, Université Paris Ouest / Rapporteur
Jean-François Dupuy
Professeur, INSA de Rennes / Examinateur
Gilles Durrieu
Professeur, Université de Bretagne Sud / Directeur de thèse
Ion Grama
Professeur, Université de Bretagne Sud / Directeur de thèse
Olivier Sire
Professeur, Université de Bretagne Sud / Examinateur

Laboratoire de Mathématiques de Bretagne Atlantique (LMBA)
UMR CNRS 6205 et Université de Bretagne-Sud
Campus de Tohannic
56000 Vannes, France

Estimation non paramètrique adaptative dans la théorie des valeurs extrêmes : application en environnement

Résumé : L'objectif de cette thèse est de développer des méthodes statistiques basées sur la théorie des valeurs extrêmes pour estimer des probabilités d'évènements rares et des quantiles extrêmes conditionnelles. Nous considérons une suite de variables aléatoires indépendantes X_{t_1}, X_{t_2}, ...,X_{t_n} associées aux temps $0 \leq t_1 < \ldots < t_n \leq T_{\max}$ où X_{t_i} a la fonction de répartition F_{t_i} et F_t est la loi conditionnelle de X sachant $T = t \in [0, T_{\max}]$. Pour chaque $t \in [0, T_{\max}]$, nous proposons un estimateur non paramétrique de quantiles extrêmes de F_t. L'idée de notre approche consiste à ajuster pour chaque $t \in [0, T_{\max}]$ la queue de la distribution F_t, par une distribution de Pareto de paramètre $\theta_{t,\tau}$ à partir d'un seuil τ. Le paramètre $\theta_{t,\tau}$ est estimé en utilisant un estimateur non paramètrique à noyau de taille de fenêtre h basé sur les observations plus grandes que τ. Sous certaines hypothèses de régularité, nous montrons que l'estimateur adaptatif proposé de $\theta_{t,\tau}$ est consistant et nous donnons sa vitesse de convergence. Nous proposons une procédure de tests séquentiels pour déterminer le seuil τ et nous obtenons le paramètre h suivant deux méthodes : la validation croisée et une approche adaptative. Nous proposons également une méthode pour choisir simultanément le seuil τ et la taille de la fenêtre h. Finalement, les procédures proposées sont étudiées sur des données simulées et sur des données réelles dans le but d'aider à la surveillance de systèmes aquatiques.

Mots clés : Estimation non paramètrique, Quantiles extrêmes conditionnelles, Probabilités d'évènements rares, Environnement.

Nonparametric adaptive estimation in the extreme value theory : application in ecology

Abstract : The objective of this PhD thesis is to develop statistical methods based on the theory of extreme values to estimate the probabilities of rare events and conditional extreme quantiles. We consider independent random variables $X_{t_1}, \cdots, X_{t_n}$ associated to a sequence of times $0 \leq t_1 < \cdots < t_n \leq T_{\max}$ where X_{t_i} has distribution function F_{t_i} and F_t is the conditional distribution of X given $T = t \in [0, T_{\max}]$. For each $t \in [0, T_{\max}]$, we propose a nonparametric adaptive estimator for extreme quantiles of F_t. The idea of our approach is to adjust the tail of the distribution function F_t with a Pareto distribution of parameter $\theta_{t,\tau}$ starting from a threshold τ. The parameter $\theta_{t,\tau}$ is estimated using a nonparametric kernel estimator of bandwidth h based on the observations larger than τ. We propose a sequence testing based procedure for the choice of the threshold τ and we determine the bandwidth h by two methods : cross validation and an adaptive procedure. Under some regularity assumptions, we prove that the adaptive estimator of $\theta_{t,\tau}$ is consistent and we determine its rate of convergence. We also propose a method to choose simultaneously the threshold τ and the bandwidth h. Finally, we study the proposed procedures by simulation and on real data set to contribute to the survey of aquatic systems.

Keywords : Nonparametric estimation, Conditional extreme quantiles, Probabilities of rare events, Environment.

Table des matières

Introduction générale

Les activités humaines génèrent et continuent de générer de multiples pollutions. Il est établi que les phénomènes météorologiques extrêmes enregistrés sur toute la planète sont étroitement liées aux pollutions et au réchauffement climatique induit par les activités humaines. Une prise de conscience se développe dans de très nombreux pays à travers le monde mais il est clair que le problème va s'aggraver à court et moyen terme si les sociétés ne réagissent pas. Dans le domaine de l'environnement, des réglementations et des contrôles sur la qualité des eaux ont été mis en place. Parmi ces techniques de contrôles, les bioindicateurs sont de plus en plus utilisés et sont très efficaces par leurs capacités à révéler la présence de concentrations très faibles de contaminants via par exemple leur accumulation dans des tissus animaux ou végétaux spécifiques, ou via des modifications sur des structures populationnelles, etc.

Dans ce travail, nous nous intéressons particulièrement au développement d'une méthode statistique d'analyse du comportement extrême de mollusques bivalves, qui permet d'aborder leur éthologie, mais qui est aussi un moyen de surveillance de la qualité de l'eau. Le cas des mollusques bivalves est particulièrement intéressant en tant qu'espèce bioindicatrice car se sont des animaux sédentaires qui peuvent être témoins de pollutions locales éventuelles et on les trouve partout dans le monde, des tropiques aux pôles. Ce travail est le fruit d'une collaboration étroite depuis plusieurs années avec l'équipe écotoxicologie aquatique de l'UMR CNRS 5805 et

de l'Université de Bordeaux. La compréhension des mécanismes fondamentaux du réglage de l'activité ventilatoire des huîtres et des bivalves en général reste un problème largement non résolu de physiologie respiratoire chez les animaux aquatiques et d'écophysiologie. Comprendre ces mécanismes est un problème d'écologie qui vise à mieux comprendre de nouveaux aspects de leur comportement trophique et de leur croissance ce qui permet de mieux gérer leurs populations, ainsi que de savoir comment elles peuvent influencer leur milieu et comment elles peuvent être influencées par ce milieu.

Le suivi du comportement de bivalves permet donc de rendre compte jour après jour de leur état de santé et, au delà, de l'évolution de la qualité de l'eau. Nous utilisons ici comme technique de surveillance du milieu aquatique la valvométrie (mesure de l'activité des valves de mollusques). Les premiers enregistrements d'activité de bivalves ont été publiés en 1909 par Marceau [1] mais depuis, la technique a été souvent reprise pour évoluer régulièrement. En très bref, il s'agit de fixer sur les valves des animaux un système d'enregistrement qui permet de suivre en permanence l'état d'ouverture/fermeture des animaux ainsi que les divers mouvements associés. Tous les systèmes utilisés à l'heure actuelle étaient basés sur la fixation d'une valve de l'animal sur le dispositif expérimental. La seconde valve était libre et portait un seul capteur. L'imagerie vidéo est apparue comme alternative intéressante mais avec les limitations qu'on imagine pour une optique de caméra devant éventuellement rester sous l'eau pour de longues périodes. Deux applications commerciales sont le Mosselmonitor [2] et le Dreissena monitor [3]. L'Institut français de recherche pour l'exploitation de la mer (Ifremer) a aussi développé un système avec la société Micrel NKE.

Au sein de l'UMR 5805 EPOC de l'Université de Bordeaux, une équipe pluridisciplinaire composée des chercheurs Jean-Charles Massabuau (DR CNRS), Damien Tran (CR 1 CNRS), de Gilles Durrieu (PR, UBS) et des ingénieurs Pierre Ciret (IR CNRS) et Mohamedou Sow (IR ADERA) a développé un système permettant de travailler non pas sur des animaux collés (cas

des systèmes précédents) mais sur des animaux libres de tous mouvements pour de longues, voir très longues (plusieurs mois) périodes de temps. Des animaux stressés par des contraintes expérimentales peuvent en effet présenter une perte de sensibilité très importante. La valvométrie a d'abord été développée en laboratoire pour la recherche fondamentale. Les buts étaient de mieux comprendre certains aspects de physiologie respiratoire et d'écophysiologie en respectant au mieux le comportement naturel spontané de petits bivalves d'eau douce, les corbicules (*Corbicula fluminea*). A partir de 2006, la technologie a évolué pour permettre le travail sur le terrain afin de caractériser et traiter statistiquement le signal obtenu sur de très longues séries. Un des buts est de faire du suivi à très long terme afin de comprendre le fonctionnement et la trajectoire temporelle des systèmes environnementaux, leurs réponses à des évènements perturbateurs et leur dynamique d'acclimatation.

Aujourd'hui la valvométrie HFNI (Haute Fréquence Non Invasive) permet d'enregistrer, pratiquement en ligne, les mouvements des valves à haute fréquence sans interférer avec le comportement normal des animaux, n'importe où dans le monde pendant des périodes de 1 à 2 ans sans intervention humaine *in-situ*. La valvométrie est ainsi devenue une technique qui permet d'enregistrer les réactions de bivalves sur le terrain, proche ou distant, face aux changements de la qualité de l'eau dans laquelle ils vivent. L'idée de base est que les mollusques bivalves ventilent tout au long de la journée pour se nourrir et respirer. Comme ils ne disposent la plupart du temps pas de moyens de fuite bien efficace, leur protection principale est basée sur la fermeture de leurs valves. De nombreuses anomalies telles que le contact d'un prédateur ou l'arrivée d'un contaminant dans la colonne d'eau, détectées (grâce aux différents types de mécano- et chimiorécepteurs) comme dangereuses pour l'animal, se traduisent par des comportements de protection allant de la fermeture brève à la fermeture prolongée en passant par des modifications de la fréquence et des vitesses des séquences d'ouverture et de fermeture des valves. Lorsque ces comportements affectent non pas uniquement un seul individu mais le

groupe, ces modifications sont le reflet de perturbations du milieu proche des animaux et donc le comportement valvaire peut être alors un excellent outil d'aide à la gestion et à la surveillance du milieu ([4–11]).

Le dispositif expérimental développé par l'UMR 5805 EPOC permet de suivre 24h/24 le comportement *in situ* de lots de 16 bivalves. Plusieurs dispositifs sont installés à différents endroits du monde, sous la jetée d'Eyrac à Arcachon, dans le lagon sud de Nouvelle Calédonie sur un lot de 16 bénitiers, sur la côte Atlantique française et espagnole (Port de Santander), en Norvège (Longyearbyen au Svalbard) et en Russie (région de Murmansk, en mer de Barents). Nous nous focalisons dans cette thèse sur le site de Locmariaquer dans le Morbihan en Bretagne Sud tout près de l'Université de Bretagne Sud. Une caractéristique particulière de ce projet est son approche, totalement pluridisciplinaire, alliant biologie, électronique et mathématiques appliquées (statistiques).

Le site de Locmariaquer est équipé depuis le 3 mars 2011 dans le cadre du projet transverse ASPEET (Approche Systémique du Pilotage Economique et Environnemental des Territoires) financé par l'université de Bretagne Sud. L'objectif de ce projet était de caractériser une "signature" comportementale d'un milieu, de la nature d'un contaminant ou d'un facteur du milieu en utilisant l'huître *Crassostrea Gigas* comme bioindicateur. Ces facteurs peuvent être soit des algues toxiques, soit des déchets de l'activité anthropique comme les métaux lourds, des pétroles, soit liés au réchauffement global. Les écartements valvaires de 16 huîtres ont été et sont mesurées 24h/24. Les données sont accumulées et stockées pendant 24h sur site (chaque fichier représente environ 18M octets pour une journée) puis transmises sur une station de travail par protocole de transfert de données GPRS (General Packet Radio Service) en utilisant le réseau de la téléphonie mobile et Internet (FTP). Cet important volume de données enregistrées à haute fréquence avec un nombre de variables pouvant être important (couplage avec d'autres capteurs comme des sondes multi-paramètres), nécessite le développement de modèles statistiques

performants afin de bien décrire le comportement des animaux *in situ* rapidement. En effet, un des buts est de ne pas laisser s'accumuler des données non traitées et d'obtenir immédiatement des métadonnées évolutives, présentées sous forme de graphiques lisibles instantanément.

Le but des méthodes statistiques développées et appliquées a été entre autres d'extraire des rythmes biologiques et de les présenter sous forme graphique simple pour permettre, par la caractérisation de perturbations de ces rythmes, de détecter *in fine* une pollution du milieu. Notre hypothèse de travail est que les organismes aquatiques doivent présenter des perturbations de leur comportement de base et de leurs rythmes biologiques face aux contraintes du milieu. Dans ce raisonnement, on pense que les caractéristiques typiques de leurs activités physiologiques ou comportementales seront modifiées et seront le témoin sensible et immédiat de cette perturbation du milieu. Nous avons utilisé plusieurs approches afin d'aborder divers aspects du comportement décrivant au niveau du jour, de la semaine, du mois ou sur une plus longue période de temps des traits de vie caractéristiques d'un mollusque bivalve.

Actuellement et en partie grâce aux résultats de nos travaux, l'acquisition, le transfert, et le traitement des données fonctionnent de manière automatique quelque soit le site où un système est positionné dans le monde. C'est ce qui est fait en routine pour les différents sites et en particulier pour Locmariaquer. Les enregistrements et les résultats du traitement statistique sont accessibles sur le site web L'oeil du mollusque (http://molluscan-eye.epoc.u-bordeaux1.fr/).

Au final, nous recherchons dans cette thèse à caractériser une signature comportementale extrême d'un milieu, de la nature d'un contaminant ou d'un facteur du milieu. L'approche étant basée sur une étude à haute fréquence, il nous faut des moyens statistiques pour décrire et approfondir l'étude des réactions valvaires en fonction des variations des caractéristiques du biotope.

L'objectif général de cette thèse est de développer des modèles statistiques originaux basés

sur la théorie des valeurs extrêmes sur des volumes de données importants afin de fournir une meilleure connaissance du fonctionnement et du comportement de l'huître creuse *Crassostrea gigas* dans son environnement naturel. Nous cherchons à caractériser/à discriminer des perturbations extrêmes de l'activité des animaux par l'étude d'un bivalve filtreur (l'huître dans le golfe du Morbihan), de la nature d'un contaminant ou d'un facteur du milieu.

Cette thèse s'articule autour de quatre chapitres. Le Chapitre I concerne le développement d'une méthode permettant d'estimer des probabilités d'événements rares et des quantiles extrêmes conditionnelles. Nous considérons X_{t_1}, X_{t_2}, ...,X_{t_n} des observations indépendantes associées aux temps $0 \leq t_1 < \ldots < t_n \leq T_{\max}$ où X_{t_i} a la fonction de répartition F_{t_i} et F_t est la loi conditionnelle de X sachant $T = t \in [0, T_{\max}]$. Pour chaque $t \in [0, T_{\max}]$, nous proposons un estimateur adaptatif non paramétrique de quantiles extrêmes de F_t. L'idée de notre approche consiste à ajuster pour chaque $t \in [0, T_{\max}]$ la queue de la distribution F_t, par une distribution de Pareto de paramètre $\theta_{t,\tau}$ à partir d'un seuil τ. Le paramètre $\theta_{t,\tau}$ est estimé en utilisant un estimateur non paramétrique à noyau de taille de fenêtre h basé sur les observations plus grandes que τ. Sous certaines hypothèses de régularité, nous montrons que l'estimateur adaptatif proposé de $\theta_{t,\tau}$ est consistant et nous donnons sa vitesse de convergence. Nous proposons une procédure de tests séquentiels pour déterminer le seuil τ et nous estimons le paramètre h par validation croisée. Le Chapitre II est dédiée à la construction d'une approche adaptative pour estimer la taille de la fenêtre h. Nous proposons également une méthode pour choisir simultanément le seuil τ et la taille de la fenêtre h. Les Chapitres III et IV présentent les propriétés de notre procédure sur des simulations et sur des données réelles dans le but d'estimer des changements globaux (pollution, changements de température) et ainsi d'aider à la surveillance de systèmes aquatiques. Enfin, nous concluons cette thèse par une conclusion générale et nous décrivons également quelques perspectives de recherche.

Chapitre I

Estimation non paramétrique des probabilités conditionnelles d'évènements rares et des quantiles extrêmes conditionnels

1 Introduction

La modélisation et la prédiction des événements extrêmes (tempête, tremblement de terre, inondation, crises financières, pollution, génétique, etc) est aujourd'hui un domaine de recherche très actif, notamment par l'importance de leurs impacts économiques, sociaux, génétiques et environnementaux. On peut noter depuis quelques années un intérêt croissant pour l'application de la théorie des valeurs extrêmes pour la modélisation de tels évènements. Pour une présentation complète du sujet, nous renvoyons à l'ouvrage de référence [12] rappelant les principaux résultats théoriques de la théorie des valeurs extrêmes et à Reiss et Thomas [13] qui proposent quelques exemples pratiques dans les domaines de la finance, de l'assurance et aussi en sciences environnementales.

Soit $F_t(x) = P(X \leq x|T = t)$ la fonction de répartition conditionnelle d'une variable aléatoire X définie sur $[x_0, \infty)$ sachant $T = t \in [0, T_{\max}]$, où nous supposons que F_t appartient au domaine d'attraction de la loi de Fréchet. Nous observons les variables aléatoires indépendantes $X_{t_1}, \ldots, X_{t_n}$ associées aux temps $0 \leq t_1 < \ldots < t_n \leq T_{\max}$ où X_{t_i} a la distribution F_{t_i}.

Dans ce Chapitre, nous commençons par présenter le modèle et les estimateurs. Pour chaque $t \in [0, T_{\max}]$, nous proposons un estimateur non paramétrique de la queue de la distribution, des

probabilités d'évènement rare et des quantiles extrêmes de F_t. Grâce au théorème de Fisher-Tippett-Gnedenko, la queue de la distribution de F_t peut être approchée par la queue d'une distribution de Pareto de paramètre $\theta_{t,\tau}$ à partir d'un seuil τ. Le paramètre $\theta_{t,\tau}$ est estimé en utilisant un estimateur non paramétrique à noyau de taille de fenêtre h basé sur les observations plus grandes que τ et nous proposons une procédure automatique ponctuelle pour déterminer le seuil τ. La taille de la fenêtre h est déterminée par la méthode de la validation croisée mais nous proposons aussi dans le Chapitre II une méthode adaptative. Sous certaines hypothèses de régularité, nous montrons que les estimateurs non adaptatif et adaptatif de $\theta_{t,\tau}$ proposés sont consistants et nous donnons leurs vitesses de convergence. Les propriétés des procédures proposées seront également étudiées par simulation dans le Chapitre III et dans le Chapitre IV sur des données réelles dans le domaine environnementale.

1.1 Modèle et estimateurs

L'idée de la méthode est de déterminer le seuil τ et d'ajuster sur $[\tau, +\infty[$ une distribution de Pareto définie par

$$G_{\tau,\theta}(x) = 1 - \left(\frac{x}{\tau}\right)^{-\frac{1}{\theta}} \qquad x \in [\tau, +\infty[,$$

où le paramètre $\theta > 0$ et $\tau \geq x_0$ est la valeur inconnue du seuil. Nous obtenons ainsi un modèle semi-paramétrique défini par

$$F_{t,\tau,\theta}(x) = \begin{cases} F_t(x) & \text{if } x < \tau \\ 1 - (1 - F_t(\tau))(1 - G_{\tau,\theta}(x)) & \text{if } x \geq \tau. \end{cases} \tag{I.1}$$

Soit $\mathcal{K}(P, Q) = \int \log \frac{dP}{dQ}\, dP$ la divergence de Kullback-Leibler entre deux mesures équivalentes P et Q. Soit $F_{t,\tau}$ la fonction de répartition d'excès au dessus du seuil τ définie par :

$$F_{t,\tau}(x) = 1 - \frac{1 - F(x)}{1 - F(\tau)}, \qquad x \geq \tau.$$

Pour chaque $t \in [0, T_{\max}]$ et $\tau \geq x_0$, le minimum de la divergence de Kullback-Leibler entre

$F_{t,\tau}$ et le modèle $G_{\tau,\theta}$ est atteint pour

$$\theta_{t,\tau} = \arg\min_{\theta\in\Theta} \mathcal{K}\left(F_{t,\tau}, G_{\tau,\theta}\right) = \int_{\tau}^{\infty} \log\frac{x}{\tau}\frac{F_t\left(dx\right)}{1-F_t\left(\tau\right)}, \tag{I.2}$$

Pour chaque t fixé dans $[0, T_{\max}]$ et pour $\tau \geq x_0$, nous construisons un estimateur non paramétrique à noyau du paramètre fonctionnel $t \to \theta_{t,\tau}$. Nous estimons dans un premier temps la fonction $\theta_{t,\tau}$ au point t en utilisant un estimateur à noyau K d'une taille de fenêtre h et dans un second temps nous donnons une procédure de sélection du seuil τ. En maximisant la quasi-log vraisemblance pondérée par rapport à θ, nous obtenons l'estimateur

$$\widehat{\theta}_{t,h,\tau} = \frac{1}{\widehat{n}_{t,h,\tau}} \sum_{X_{t_i}>\tau} W_{t,h}\left(t_i\right) \log\left(\frac{X_{t_i}}{\tau}\right), \tag{I.3}$$

où $W_{t,h}\left(t_i\right) = K\left(\frac{t_i - t}{h}\right)$ avec K une fonction noyau et $\widehat{n}_{t,h,\tau} = \sum_{i=1}^{n} W_{t,h}(t_i) 1_{\left\{X_{t_i}>\tau\right\}}$. L'estimateur semi-paramétrique de la fonction de répartition F_t est donné par

$$\widehat{F}_{t,h,\tau}\left(x\right) = \begin{cases} \widehat{F}_{t,h}\left(x\right), & x \in [x_0, \tau], \\ 1 - \left(1 - \widehat{F}_{t,h}\left(\tau\right)\right)\left(\frac{x}{\tau}\right)^{-\frac{1}{\widehat{\theta}_{t,h,\tau}}}, & x > \tau, \end{cases}$$

où

$$\widehat{F}_{t,h}\left(x\right) = \frac{1}{\sum_{j=1}^{n} W_{t,h}(t_j)} \sum_{i=1}^{n} W_{t,h}\left(t_i\right) 1_{\{X_{t_i}\leq x\}}$$

est la fonction de répartition empirique pondérée. L'estimateur semi-paramétrique du quantile d'ordre p est donné par

$$\widehat{q}_p\left(t\right) = \begin{cases} \widehat{F}_{t,h}^{-1}\left(p\right) & \text{pour} \quad p < \widehat{p}_0, \\ \tau\left(\frac{1-\widehat{p}_0}{1-p}\right)^{\widehat{\theta}_{t,h,\tau}} & \text{sinon}, \end{cases} \tag{I.4}$$

avec $\widehat{p}_0 = \widehat{F}_{t,h}\left(\tau\right)$.

La principale difficulté concerne les choix du seuil τ et de la taille de la fenêtre h. Dans les paragraphes qui suivent, nous donnons des procédures pour déterminer τ et h.

1.2 Choix du seuil τ

L'estimateur $\widehat{\theta}_{t,h,\tau}$ est très sensible aux choix du seuil τ et de la taille de fenêtre h. La difficulté est de choisir τ assez petit de façon à ce que l'estimateur de la fonction de répartition empirique pondérée dans le modèle (I.1) dispose de suffisamment d'observations pour assurer un bon ajustement de la queue de la distribution F_t. Par ailleurs, τ doit être aussi choisi assez grand de façon à éviter un biais d'estimation due à un mauvais ajustement de la queue de distribution. Nous proposons d'estimer le paramètre τ en utilisant une procédure séquentielle de tests d'adéquations similaire à celle proposée par Grama et Spokoiny [14, 15] en 2007 et 2008. Dans un premier temps, nous testons $\mathcal{H}_0(\tau)$ l'hypothèse nulle stipulant que F_t est le modèle semi-paramétrique de Pareto défini par (I.1).

Nous considérons $s_1, ..., s_m$ une suite de seuils triés par ordre décroissant de sorte que $s_1 \geq ... \geq s_m$ avec m fixé. Nous cherchons à déterminer un seuil s_k à partir des statistiques d'ordre $Y_1 \geq \ldots \geq Y_{M_{t,h}}$ associés à $\mathcal{Y}_{t,h} = \left\{X_{t_i} : \frac{t_i - t}{h} \in \text{supp}K\right\}$, où $\text{supp}K$ est le support du noyau K et $M_{t,h} = \text{card}(\mathcal{Y}_{t,h})$. Nous considérons une suite de tests d'adéquation en déterminant le premier seuil s_k notée $s_{\widehat{k}}$ pour lequel $\mathcal{H}_0\left(s_{\widehat{k}}\right)$ est rejetée en faveur de l'hypothèse alternative $\mathcal{H}_1(\tau)$: "F_t est le modèle semi-paramétrique de Pareto avec un point de rupture" donné par :

$$F_{t,\mu,s,\theta,\tau}(x) = \begin{cases} F_t(x) & \text{si } \; x \in [x_0, s], \\ 1 - (1 - F_t(s))(1 - G_{s,\mu}(x)) & \text{si } \; x \in (s, \tau], \\ 1 - (1 - F_t(s))(1 - G_{s,\mu}(\tau))(1 - G_{\tau,\theta}(x)) & \text{si } \; x \in (\tau, \infty), \end{cases}$$

où $\mu, \theta > 0$ et $x_0 \leq s \leq \tau$. La statistique de test utilisée est le maximum du rapport de log vraisemblance entre deux modèles $\mathbf{Z}_h(s_k)$. L'hypothèse nulle $\mathcal{H}_0(s_k)$ est rejeté si $\mathbf{Z}_h(s_k) > D$, où D est la valeur critique. Cette dernière est calculée par la méthode de Monte Carlo. Dans un deuxième temps, nous sélectionnons le modèle qui maximise par rapport à $\tau \in \mathcal{Y}_{t,h}$ la fonction de vraisemblance pénalisée donnée par :

$$\mathcal{L}_{t,h}\left(\tau, \widehat{\theta}_{t,h,\tau}\right) - \text{Pen}_{t,h}\left(\tau, \widehat{\theta}_{t,h,\widehat{s}}\right) = \widehat{n}_{t,h,\tau}\mathcal{K}\left(\widehat{\theta}_{t,h,\tau}, \widehat{\theta}_{t,h,\widehat{s}}\right), \tag{I.5}$$

où $\widehat{s} = s_{\widehat{k}-1}$ et

$$\mathrm{Pen}_{t,h}(\tau,\theta) = \mathcal{L}_{t,h}(\tau,\theta), \qquad \mathcal{L}_{t,h}(\tau,\theta) = \sum_{i=1}^{n} W_{t,h}(t_i) \log \frac{dF_{t,\tau,\theta}}{dx}(X_{t_i}).$$

Nous notons dans la suite $\widehat{\tau}_{t,h}$ le seuil ainsi obtenu. Nous remarquons que la pénalisation est choisie de façon à ce que la différence (I.5) coincide à une constante près avec l'entropie de Kullback-Leibler $\mathcal{K}\left(\widehat{\theta}_{t,h,\tau}, \widehat{\theta}_{t,h,\widehat{s}}\right)$. Lorsque τ s'approche de $\widehat{s}$, cette différence tend vers 0.

1.3 Propriétés asymptotiques

Nous notons $\theta_{t,\tau}$ le paramètre définie par (I.2). Supposons que des suites τ_n et h_n vérifient la condition suivante

$$\sum_{i=1}^{n} W_{t,h_n}(t_i)\chi^2(F_{t_i}, F_{t_i,\tau_n,\theta_{t,\tau_n}}) = O(\log n) \qquad \text{quand} \qquad n \to \infty, \tag{I.6}$$

où $\chi^2(P,Q) = \int \frac{dP}{dQ} dP - 1$ est la divergence de χ^2 entre deux lois équivalentes P et Q.

Théorème I.1. *Sous la conditions (I.6) et* $\bar{n}_{t,h_n,\tau_n} = \sum_{i=1}^{n} W_{t,h_n}(t_i)(1-F_{t_i}(\tau_n)) \to \infty$ *quand* $n \to \infty$, *nous avons*

$$\mathcal{K}\left(\widehat{\theta}_{t,h_n,\widehat{\tau}_{t,h_n}}, \theta_{t,\tau_n}\right) = O_P\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right) \qquad \textit{quand} \qquad n \to \infty.$$

On déduit du Théorème I.1 que,

$$\mathcal{K}\left(F_{t,\tau_n}, G_{\tau_n,\widehat{\theta}_{t,h_n,\widehat{\tau}_{t,h_n}}}\right) = O_P\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right) \qquad \text{quand} \qquad n \to \infty.$$

Nous montrons ainsi que la vitesse de convergence obtenue est quasi-optimale.

Dans le cas du modèle de Hall [16], il existe des constantes positives $c_{\min}$, $c_{\max}$, $\gamma_{\min}$, $\gamma_{\max}$, $A_{\max} > 0$ et $\rho > 0$ telles que, pour chaque $t \in [0, T_{\max}]$, la famille $(F_t)_{t\in(0,T_{\max}]}$ satisfait $F_t(x_0) = 0$ et

$$f_t(x) = \frac{c_t}{\gamma_t} x^{-\frac{1}{\gamma_t}-1}(1 + r_t(x)), \quad |r_t(x)| \le A_t x^{-\frac{\rho}{\gamma_t}} \text{ quand } x \to \infty, \tag{I.7}$$

où γ_t, c_t et A_t sont des fonctions dépendantes du temps vérifiant $c_{\min} \leq c_t \leq c_{\max}$, $\gamma_{\min} \leq \gamma_t \leq \gamma_{\max}$, $A_t \leq A_{\max}$.

Théorème I.2. *Sous des conditions de régularités, nous obtenons les vitesses de convergence pour le modèle I.7 :*

$$\sqrt{\mathcal{K}\left(\widehat{\theta}_{t,h_n,\widehat{\tau}_n}, \theta_{t,\tau_n}\right)} = O_{\mathbf{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{1+\beta(2+1/\rho)}}\right) \qquad quand \qquad n \to \infty,$$

et

$$\sqrt{\mathcal{K}\left(F_{t,\tau_n}, G_{\tau_n,\widehat{\theta}_{t,h_n,\widehat{\tau}_{t,h_n}}}\right)} = O_{\mathbf{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{1+\beta(2+1/\rho)}}\right) \qquad quand \qquad n \to \infty,$$

où γ_t est une fonction β-Höldérienne avec $0 < \beta \leq 1$.

Le modèle de Hall englobe une grande variété de modèles comme les lois stables (non normales) et la distribution de Fréchet. Si $\rho \to \infty$, ce modèle correspond au modèle de Pareto et la vitesse devient $\left(\frac{\log n}{n}\right)^{\frac{\beta}{1+2\beta}}$ qui correspond à la vitesse de convergence d'un estimateur d'une fonction non paramétrique de paramètre de régularité β.

De la même manière, nous pouvons établir les vitesses de convergence pour le modèle de mélange de 2 distributions de Pareto.

1.4 Estimation de la taille de la fenêtre h par la méthode de la validation croisée

Le choix du paramètre h est un point crucial. Nous proposons une méthode basée sur une approche de type validation croisée. Nous considérons $\mathcal{H} = \{h_m : h_m = h_0 q^m, m = 1, \ldots, M_h\}$ avec $q > 1$, $h_0 > 0$ et M_h grand. Nous proposons la fonction de validation croisée :

$$CV(h_m, p) = \frac{1}{M_h card(T_{grid})} \sum_{h_l \in \mathcal{H}} \sum_{t_i \in T_{grid}} \psi\left(\widehat{F}^{-1}_{t_i,h_l}(p), \widehat{q}_p^{(-i)}(t_i, h_m)\right), \tag{I.8}$$

où T_{grid} est une suite de points d'une grille régulière sur $[0, T_{\max}]$, $\widehat{q}_p^{(-i)}(t_i, h_m)$ désigne un estimateur du quantile extrême d'ordre p au point t_i défini par (I.4) calculé sur l'échantillon

privé de l'observation X_{t_i} et $\psi(x, y) = |\log x - \log y|$, $x, y > 0$. L'estimateur de h notée h_{CV} s'obtient par minimisation par rapport à h_m de la fonction $CV(h_m, p)$ pour p fixé.

1.5 Simulation et application

Une étude par simulation est donnée dans notre article soumis et nous détaillons davantage le comportement de notre approche sur des simulations dans le Chapitre III et sur des données réelles environnementales dans le Chapitre IV dans un objectif de surveillance de la qualité des eaux.

2 Article 1 : Nonparametric adaptive estimation of conditional probabilities of rares events and extreme quantiles

Ce paragraphe concerne un article soumis pour publication dans Extremes. Cet article de 30 pages a été écrit par Gilles Durrieu, Ion Grama, Quang-Khoai Pham et Jean-Marie Tricot.

2.1 Introduction

Extreme values modeling and estimation play an important role in many practical applications in different areas of biology, medicine, environment, finance etc. We are interested in recovering the tail of the conditional distribution $F_t(x) = P(X \leq x|T = t)$ given the time t from independent random variables $X_{t_1}, \ldots, X_{t_n}$ with distribution functions $F_{t_1}, \ldots, F_{t_n}$, in the context of the fixed design points $0 \leq t_1 < \ldots < t_n \leq T_{\max} < \infty$. In this article, we propose a statistical method to estimate the tail conditional probabilities F_t and extreme p-quantiles $F_t^{-1}(p)$.

There is a vast literature on this problem. Smith [17] and Davidson and Smith [18] considered a parametric regression model for conditional extremes. Hall and Tajvidi [19] and Davison and Ramesh [20] studied semiparametric models with a parametric structure for the distribution function F_t in conjunction with a local nonparametric estimation in time. Beirlant and Goegebeur [21] analyzed a semiparametric model based on an exponential regression for log-spacings of the generalized residuals whereas Beirlant and Goegebeur [22] considered local polynomial estimators for the tail index using generalized regression linear model with multidimensional covariates. Gardes and Girard [23] developed an estimator of the conditional tail index based on a weighted sum of the log-spacings between observations selected in a moving window. Gardes and Girard [24] estimated the conditional extreme quantile by nearest neighbor approach. In random design, Goegebeur et al. [25], Gardes and Stupfler [26] and Stupfler [27] proposed non

parametric estimation of the conditional tail index in a regression context using respectively weighted sums of power transformation, smoothed local Hill and moment estimators.

Our estimation is based on adjusting a Pareto distribution function for observations beyond a given threshold τ in a neighborhood $[t-h, t+h]$ of the time t. It is known since Embrechts, Klüppelberg and Mikosh [12] that one of the difficulties in the statistics of extremes is the choice of the threshold τ which determines the number of upper statistics used in the estimation. In most of the papers mentioned above, the threshold τ is fixed and the focus is on the choice of the bandwidth h. However, if the number of upper statistics is too small the estimator exhibits large variability and if it is too high the bias may be important. Even in the i.i.d. setting, the adaptive choice of τ is a challenging problem : we address to Hall and Welsh [28], Dress and Kaufmann [29], Guillou and Hall [30], Huisman et al. [31], Beirlant et al. [32], Grama and Spokoiny [14, 15] and El Methni et al. [33]. The appropriate choice of τ is especially important when the distribution functions F_t do not belong to a parametric family of distributions. In this paper we give a data driven threshold τ to adjust the tails with a Pareto model from data in the neighborhood $[t-h, t+h]$ at each time t. The choice of the bandwidth parameter h is given globally based on a cross-validation procedure.

We firstly determine the rate of convergence of the estimators of the adjusted model parameters when the bandwidth and the threshold are deterministic. To illustrate our more general results, we consider the particular case of the family of time varying Hall models defined by the densities $f_t(x) = \frac{c_t}{\gamma_t} x^{-\frac{1}{\gamma_t}-1}\left(1 + r_t\left(x\right)\right)$, where $|r_t(x)| \leq A_t x^{-\frac{\rho}{\gamma_t}}$ as $x \to \infty$ and γ_t, c_t, A_t are some time depending functions satisfying Hölder condition of order β (see Section 2.3.2 for more details). We obtain a rate of convergence of order $(n^{-1}\log n)^{\frac{\beta}{1+\beta(2+1/\rho)}}$ as $n \to \infty$. Note that when ρ goes to ∞ the distribution F_t becomes parametric : the rate of convergence is then $(n^{-1}\log n)^{\frac{\beta}{1+2\beta}}$ which up to a logarithmic multiple coincides the optimal rate given by Stone [34] and Gardes and Girard [23] (see their Corollary 2). The advantage of the proposed model

is that it covers more general models where the distribution function F_t can be nonparametric.

Secondly, we give a selection procedure for the choice of the threshold τ and we construct an adaptive estimator. Our selection procedure consists of goodness-of-fit multiple tests for the parametric-based part of the model. Thereby we validate the adjusted tail : at each step of the procedure the adjusted tail is tested and if it is not rejected the support of the parametric model is enlarged and tested again until the parametric model is rejected. Moreover we justify theoretically the obtained adaptive estimator : the convergence results established for the deterministic threshold are extended to the adaptive estimator. We study the developed adaptive procedure by simulations.

We apply the developed method to data collected in an ecological study. Our application concerns the activity of oysters. The ability of oyster to permanently "taste" their environment is one of the possible ways to monitor the quality of the coastal waters and read throughout the year the health of both the oysters and their environment. Ecological studies often focus on average and median effects of environmental factors such as temperature, precipitation, salinity etc., but ecological dynamic is strongly affected by environmental extremes events, as seen in Denny et al. [35]. Extreme events can cause dramatic ecological change whose recovery often is not possible. Such effects arise when populations are pushed below some minimum density threshold (e.g., the Allee effect) or when a community or ecosystem enters an alternate stable state (Allee et al. [36] and Folke et al. [37]). For instance, mortalities of Pacific oysters during the summer months have been documented throughout the world and can affect heavily the juveniles and matures (Samain and McCombie [38]), with extreme cases involving more than 90% mortality (Burge et al. [39]). The aim of our experiments is to propose a water quality monitoring system through the observation of extreme oyster's behavior.

The paper is organized as follows. Section 2.2 describes the model and estimators. The asymptotic results of the estimators are stated in Section 2.3, with proofs given in the Section

2.4. In Section 2.4 we construct our adaptive estimator and prove its convergence. In Section 2.5 we report simulation results to study the performance of the proposed procedure. Section 2.6 presents an application to ecological data sets. Finally, we conclude with a discussion in Section 2.7.

2.2 Model and Estimator

We consider a pair of random variables (X, T), where X represents a quantity of interest and $T \in [0, T_{\max}]$ the time. Let $F_t(x) = P(X \leq x | T = t)$ be the conditional distribution of X given $T = t$ supported on the interval $[x_0, \infty)$, $x_0 \geq 0$ with a strictly positive density f_t. We assume that the distributions F_t are in the domain of attraction of the Fréchet distribution. We observe independent random variables $X_{t_1}, ..., X_{t_n}$ associated to a sequence of times $0 \leq t_1 < ... < t_n \leq T_{\max}$ where for each t_i the random variable X_{t_i} has a distribution function F_{t_i}. Given $x > x_0$ and $p \in (0, 1)$, the main aim is to provide a pointwise estimate of the tail probability $S_t(x) = 1 - F_t(x)$ and the extreme p-quantile $F_t^{-1}(p)$ processes on $[0, T_{\max}]$.

The empirical survival function is routinely used to estimate $S_t(x)$, but this estimator does not provide a reliable estimation for large values of x, due to the lack of observations in this range. Otherwise, a parametric model is fitted to data and the values $S_t(x)$ and $F_t^{-1}(p)$ are inferred from the corresponding local fit. However this may cause a severe bias if the fitted model is misspecified. We combine the flexible empirical distribution function and the parametric fit in one model. The idea is to adjust, for some $\tau \geq x_0$, the excess distribution function

$$F_{t,\tau}(x) = 1 - \frac{1 - F_t(x)}{1 - F_t(\tau)}, \quad x \in [\tau, \infty) \tag{I.9}$$

by a parametric model which has good prediction properties. Here, motivated by the Fisher-Tippett-Gnedenko theorem ([32], Theorem 2.1 page 75), we choose a Pareto distribution defined

by

$$G_{\tau,\theta}(x) = 1 - \left(\frac{x}{\tau}\right)^{-\frac{1}{\theta}}, \quad x \in [\tau, \infty), \tag{I.10}$$

where $\theta > 0$ is a parameter and $\tau \geq x_0$ is a threshold value.

It is possible to consider the generalized Pareto distribution (Berlant and Goegebeur [22], Diebolt et al. [40], Carreau and Girard [41], among others) but in this article we focus on a standard Pareto distribution given by (I.10). The results given in our simulation studies and the application show that the standard Pareto provides excellent approximation in the Frechet domain of attraction and remains robust considering different models.

On the interval $[x_0, \tau]$, we estimate F_t by the empirical distribution function, while beyond τ we use the adjusted probability $G_{\tau,\theta}$ where θ has to be estimated. So, we consider the semi-parametric model defined by

$$F_{t,\tau,\theta}(x) = \begin{cases} F_t(x) & \text{if } x \in [x_0, \tau], \\ 1 - (1 - F_t(\tau))(1 - G_{\tau,\theta}(x)) & \text{if } x > \tau. \end{cases} \tag{I.11}$$

We study the asymptotic properties of the weighted maximum quasi-likelihood estimator of θ and propose a selection rule to determine the threshold τ for a given value of t.

Let $K(\cdot)$ be a kernel function assumed to be continuous, non-negative, symmetric with support on the real line such that $K(x) \leq 1$ and define the weights $W_{t,h}(t_i) = K\left(\frac{t_i - t}{h}\right)$, where $h > 0$ is a bandwidth parameter. For any $t \in [0, T_{\max}]$, the weighted quasi-log-likelihood function (see Staniswalis [42], Loader [43]) is

$$\begin{aligned} \mathcal{L}_{t,h}(\tau, \theta) &= \sum_{i=1}^{n} W_{t,h}(t_i) \log \frac{dF_{t,\tau,\theta}}{dx}(X_{t_i}) \\ &= \sum_{i=1}^{n} 1_{\{X_{t_i} \leq \tau\}} W_{t,h}(t_i) \log f_t(X_{t_i}) \\ &\quad + \sum_{i=1}^{n} 1_{\{X_{t_i} > \tau\}} W_{t,h}(t_i) \log \left(\frac{(1 - F_t(\tau))}{\tau\theta} \left(\frac{X_{t_i}}{\tau}\right)^{-\frac{1}{\theta}-1} \right), \end{aligned}$$

where 1_A takes the values 1 when the condition A is verified and 0 otherwise. Maximizing

$\mathcal{L}_{t,h}(\tau,\theta)$ with respect to θ, we obtain the estimator

$$\widehat{\theta}_{t,h,\tau} = \frac{1}{\widehat{n}_{t,h,\tau}} \sum_{i=1}^{n} W_{t,h}(t_i) 1_{\{X_{t_i}>\tau\}} \log\left(\frac{X_{t_i}}{\tau}\right) \tag{I.12}$$

where

$$\widehat{n}_{t,h,\tau} = \sum_{i=1}^{n} W_{t,h}(t_i) 1_{\{X_{t_i}>\tau\}} \tag{I.13}$$

is the weighted number of the observations beyond the threshold τ. The distribution function $F_t(x)$ at time t is then estimated by

$$\widehat{F}_{t,h,\tau}(x) = \begin{cases} \widehat{F}_{t,h}(x) & \text{if } x \in [x_0,\tau], \\ 1 - \left(1-\widehat{F}_{t,h}(\tau)\right)\left(\frac{x}{\tau}\right)^{-1/\widehat{\theta}_{t,h,\tau}} & \text{if } x > \tau, \end{cases} \tag{I.14}$$

which combines the weighted empirical distribution function

$$\widehat{F}_{t,h}(x) = \frac{1}{\sum_{j=1}^{n} W_{t,h}(t_j)} \sum_{i=1}^{n} W_{t,h}(t_i) 1_{\{X_{t_i}\le x\}}$$

and the fitted Pareto law. For any $p \in (0,1)$, the estimator of the p-quantile of X_t is defined by

$$\widehat{q}_p(t) = \widehat{q}_p(t,h) \equiv \begin{cases} \widehat{F}_{t,h}^{-1}(p) & \text{if } \quad p < \widehat{p}_\tau, \\ \tau\left(\frac{1-\widehat{p}_\tau}{1-p}\right)^{\widehat{\theta}_{t,h,\tau}} & \text{otherwise,} \end{cases} \tag{I.15}$$

where $\widehat{p}_\tau = \widehat{F}_{t,h}(\tau)$.

2.3 Asymptotic properties

2.3.1 Main results

Let $\mathcal{K}(P,Q) = \int \log \frac{dP}{dQ} dP$ be the Kullback-Leibler entropy between two equivalent measures P and Q. For the Kullback-Leibler entropy between two Pareto distributions $G_{\tau,\theta'}$ and $G_{\tau,\theta}$ we use the notation :

$$\mathcal{K}(\theta',\theta) = \mathcal{K}(G_{\tau,\theta'}, G_{\tau,\theta}) = G\left(\frac{\theta'}{\theta} - 1\right)$$

where

$$G(x) = x - \log(x+1), \quad x > -1.$$

The χ^2 entropy between P and Q is defined by $\chi^2(P,Q) = \int \frac{dP}{dQ} dP - 1$. By Jensen's inequality we have $\chi^2(P,Q) \geq 0$. For any non-negative random variables A_n and B_n, the notation $A_n = O_{\mathbf{P}}(B_n)$ as $n \to \infty$ means that there exists a constant $c > 0$ such that $\mathbf{P}(A_n \leq cB_n) \to 1$ as $n \to \infty$.

For any $t \in [0, T_{\max}]$ denote

$$\overline{n}_{t,h,\tau} = \sum_{i=1}^{n} W_{t,h}(t_i)(1 - F_{t_i}(\tau)). \tag{I.16}$$

The number $\overline{n}_{t,h,\tau}$ can be interpreted as the mean number of weighted observations exceeding the threshold τ associated to t. If the kernel K has a finite support $\mathrm{supp}K$ then $\overline{n}_{t,h,\tau}$ is the mean number of weighted observations X_{t_i} exceeding the threshold τ with $t_i \in \mathrm{supp}K$.

The main result of the paper is the following theorem which provides an oracle inequality for the estimator $\widehat{\theta}_{t,h,\tau}$.

Theorem I.1. *Assume that* $\{\tau_n\}$ *and* $\{h_n\}$ *are two sequences such that* $\tau_n \geq x_0$ *and*

$$\overline{n}_{t,h_n,\tau_n} \to \infty \quad as \quad n \to \infty. \tag{I.17}$$

Then, for any sequence of positive numbers $\{\theta_n\}$, *we have as* $n \to \infty$,

$$\mathcal{K}\left(\widehat{\theta}_{t,h_n,\tau_n}, \theta_n\right) = O_{\mathbf{P}}\left(\frac{\log n}{\overline{n}_{t,h_n,\tau_n}} + \frac{1}{\overline{n}_{t,h_n,\tau_n}} \sum_{i=1}^{n} W_{t,h_n}(t_i)\chi^2(F_{t_i}, F_{t_i,\tau_n,\theta_n})\right). \tag{I.18}$$

Proof. See Section 2.8.1. □

The first term in the bound (I.18) is referred as the stochastic error while the second one is the weighted square modelling bias induced by the use of the local parametric tail instead of the true one in the neighborhood of the estimation point t. In the case when K has the compact

support $[-1,1]$, the bias term can be bounded as follows :

$$\frac{1}{\overline{n}_{t,h_n,\tau_n}}\sum_{i=1}^{n} W_{t,h_n}(t_i)\chi^2(F_{t_i},F_{t_i,\tau_n,\theta_n}) \leq \sup_{s\in[t-h_n,t+h_n]} \chi^2(F_{s,\tau_n},G_{\tau_n,\theta_n}).$$

To ensure that the second term in the right hand side of (I.18) is at least of the same order as the first one, we shall assume that the family $(F_t)_{t\in[0,T_{\max}]}$ satisfies the following small modeling bias condition (cf. [44]) :

C1. For any $t \in [0,T_{\max}]$ there exist sequences $\{\theta_n\}$, $\{\tau_n\}$ and $\{h_n\}$ (generally depending on t) such that

$$\sum_{i=1}^{n} W_{t,h_n}(t_i)\chi^2(F_{t_i},F_{t_i,\tau_n,\theta_n}) = O\left(\log n\right) \quad \text{as} \quad n\to\infty. \tag{I.19}$$

The class of distributions $(F_t)_{t\in[0,T_{\max}]}$ satisfying (I.17) and condition C1 is very large. For instance, this is the case when $(F_t)_{t\in[0,T_{\max}]}$ is a time varying mixture of Pareto models or the time varying Hall model which generalizes the model studied in [16] and [45]. Note that the Hall model includes the class of stable non-normal distributions.

The best approximation order in (I.18) is attained when the sequences $\{\theta_n\}$, $\{\tau_n\}$ and $\{h_n\}$ are chosen such that

$$\sum_{i=1}^{n} W_{t,h_n}(t_i)\chi^2(F_{t_i},F_{t_i,\tau_n,\theta_n}) \asymp \log n, \tag{I.20}$$

where $a_n \asymp b_n$ means $0 < c_1 \leq \frac{a_n}{b_n} \leq c_2 < \infty$, for any n and some constants c_1 and c_2. If (I.20) is satisfied, we say that $\{\theta_n\}$ (respectively $\{\tau_n\}$ and $\{h_n\}$) is the oracle parameter (respectively oracle threshold and oracle bandwidth). We shall see in the next section that under some regularity assumptions, (I.20) is true for $\theta_n = \theta_{t,\tau_n}$, where

$$\theta_{t,\tau} = \arg\min_{\theta>0} \mathcal{K}(F_{t,\tau},G_{\tau,\theta}) \tag{I.21}$$

is the best fitted Pareto parameter given by

$$\theta_{t,\tau} = \int_{\tau}^{\infty} \log\frac{x}{\tau}\frac{f_t(x)dx}{1-F_t(\tau)}, \tau \geq x_0. \tag{I.22}$$

From Theorem I.1, we have :

Theorem I.2. *Assume that the family* $(F_t)_{t\in[0,T_{\max}]}$ *satisfies condition C1 and*

$$\overline{n}_{t,h_n,\tau_n} \to \infty \quad as \quad n \to \infty.$$

Then, we have,

$$\mathcal{K}\left(\widehat{\theta}_{t,h_n,\tau_n}, \theta_n\right) = O_{\mathbf{P}}\left(\frac{\log n}{\overline{n}_{t,h_n,\tau_n}}\right) \quad as \quad n \to \infty.$$

Proof. This is a consequence of Theorem I.1 and condition C1. □

In the next section, we determine the explicit rate of convergence for these two models.

2.3.2 Time varying Hall model

The family $(F_t)_{t\in(0,T_{\max}]}$ is a time varying Hall model if there exists positive finite constants $c_{\min}$, $c_{\max}$, $\gamma_{\min}$, $\gamma_{\max}$, $A_{\max} > 0$ and $\rho > 0$ such that, for each $t \in [0, T_{\max}]$, the distribution function F_t satisfies $F_t(x_0) = 0$ and

$$f_t(x) = \frac{c_t}{\gamma_t} x^{-\frac{1}{\gamma_t}-1}\left(1 + r_t(x)\right), \quad |r_t(x)| \le A_t x^{-\frac{\rho}{\gamma_t}} \text{ as } x \to \infty, \tag{I.23}$$

where γ_t, c_t and A_t are some time depending functions satisfying $c_{\min} \le c_t \le c_{\max}$, $\gamma_{\min} \le \gamma_t \le \gamma_{\max}$, $A_t \le A_{\max}$.

For simplicity, we shall assume in this section that the kernel function K has the compact support $[-1, 1]$.

Proposition I.3. *Assume the time varying Hall model given by (I.23). Let* $\theta_{t,\tau}$ *be the best fitted Pareto parameter defined by (II.16). Suppose that there exist constants* $0 < \beta \le 1$ *and* $L > 0$*, such that for any* $0 \le t, s \le T$*,*

$$|\gamma_t - \gamma_s| \le L|t - s|^{\beta}. \tag{I.24}$$

Then the family $(F_t)_{t\in[0,T_{\max}]}$ *verifies condition C1 with*

$$h_n \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{1+\beta(2+1/\rho)}}, \quad \tau_n \asymp \left(\frac{n}{\log n}\right)^{\frac{\gamma_t\beta/\rho}{1+\beta(2+1/\rho)}} \quad and \quad \theta_n = \theta_{t,\tau_n}.$$

Proof. See Section 2.8.2. □

The following theorem gives an explicit rate of convergence of the estimator $\widehat{\theta}_{t,h_n,\tau_n}$.

Theorem I.4. *Under the assumptions of Proposition I.3, we have*

$$\sqrt{\mathcal{K}\left(\widehat{\theta}_{t,h_n,\tau_n},\theta_{t,\tau_n}\right)} = O_{\mathbf{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{1+\beta(2+1/\rho)}}\right) \quad as \quad n\to\infty.$$

Proof. This Theorem is a consequence of Proposition I.3 and Theorem I.2. □

We would like to comment on the obtained rate of convergence. Note that when ρ is ∞, the distribution function F_t becomes fully parametric (Pareto distribution). In this case, the rate of convergence $n^{-\frac{\beta}{2\beta+1}}$ is known to be optimal under the Lipschitz condition (see [23] and [34]). This shows that the rates of convergence provided by Theorems I.1 and I.4 are optimal up to a $\log n$ factor.

2.3.3 Mixture of two Pareto distributions

We consider that F_t is a mixture of two Pareto distributions defined by

$$F_t(x) = C(1-x^{-1/\gamma_t}) + (1-C)\left(1-x^{-1/\delta_t}\right),\ x\geq 1, \tag{I.25}$$

where $\delta_{\min}\leq\delta_t<\gamma_t\leq\gamma_{\max}$ and $C\in(0,1)$.

As in the previous section, we shall assume that the kernel function K has the compact support $[-1,1]$.

Proposition I.5. *Assume the time varying mixture of two Pareto distributions given by (I.25). Suppose that there exist constants $\beta\in(0,1]$ and $L>0$, such that for any $0\leq t,s\leq T$,*

$$|\gamma_t-\gamma_s|\leq L|t-s|^\beta,$$

and

$$|\delta_t-\delta_s|\leq L|t-s|^\beta.$$

Then the family $(F_t)_{t\in[0,T_{\max}]}$ verifies condition C1 with

$$h_n \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{1+\beta(2+1/\rho_t)}},\quad \tau_n\asymp\left(\frac{n}{\log n}\right)^{\frac{\gamma_t\beta/\rho_t}{1+\beta(2+1/\rho_t)}} \tag{I.26}$$

and

$$\theta_n = \theta_{t,\tau_n} = \frac{\gamma_t C \tau_n^{-1/\gamma_t} + \delta_t (1-C) \tau_n^{-1/\delta_t}}{C\tau_n^{-1/\gamma_t} + (1-C)\tau_n^{-1/\delta_t}},$$

where $\rho_t = \frac{\gamma_t}{\delta_t} - 1 > 0$.

Proof. See Section 2.8.3. □

The next theorem gives the explicit rate of convergence of the estimator $\widehat{\theta}_{t,h_n,\tau_n}$.

Theorem I.6. *Under the assumptions of Proposition I.5, we have*

$$\sqrt{\mathcal{K}\left(\widehat{\theta}_{t,h_n,\tau_n}, \theta_{t,\tau_n}\right)} = O_{\mathbf{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{1+\beta(2+1/\rho_t)}}\right) \quad as \quad n \to \infty.$$

Proof. This Theorem is a consequence of Proposition I.5 and Theorem I.2 . □

As in Section 2.3.2, the rate of convergence provided by Theorem I.6 is optimal up to a $\log n$ factor.

2.4 Automatic selection of the threshold τ

2.4.1 Maximum likelihood propagation procedure

An important problem concerns the choice of the threshold τ and of the bandwidth h. In this section we propose a selection procedure to determine τ. The idea of the procedure has similarities with the propagation approach proposed in [44]. In the first step of our procedure a sequence of likelihood ratio test is used to detect the maximal length local parametric fit for the tail. As soon as it is detected the next step consists in maximizing the penalized likelihood. This second part is different from the approach in [44] and is inspired by [15, 46–48]. The selection procedure stated above provides a choice of the threshold τ for each deterministic bandwidth h. As to the choice of the bandwidth h, in simulations presented in Section 2.5 we give a global selection of this parameter by cross validation.

Let $Y_1 \geq ... \geq Y_{M_{t,h}}$ be the order statistics pertaining to the observations $\mathcal{Y}_{t,h} = \left\{X_{t_i} : \frac{t_i - t}{h} \in \text{supp}K\right\}$, where supp$K$ is the support of the kernel K and $M_{t,h} = \text{card}(\mathcal{Y}_{t,h})$. We choose the threshold τ

in the set $\mathcal{Y}_{t,h}$ by maximizing the quasi-log-likelihood function

$$\max_{\theta} \mathcal{L}_{t,h}(\tau,\theta) - \mathrm{Pen}_{t,h}\left(\tau,\widehat{\theta}_{t,h,\widehat{s}}\right) = \mathcal{L}_{t,h}\left(\tau,\widehat{\theta}_{t,h,\tau}\right) - \mathrm{Pen}_{t,h}\left(\tau,\widehat{\theta}_{t,h,\widehat{s}}\right),$$

where the penalty function is defined by

$$\mathrm{Pen}_{t,h}(\tau,\theta) = \mathcal{L}_{t,h}(\tau,\theta) \tag{I.27}$$

and $\widehat{s}$ is a break time to be determined from a multiple goodness-of-fit testing procedure below.

Consider the null hypothesis $\mathcal{H}_0(\tau) : F_t = F_{t,\tau,\theta}$ where the distribution function $F_{t,\tau,\theta}$ is defined by (I.11) and the alternative hypothesis $\mathcal{H}_1(s,\tau) : F_t = F_{t,\mu,s,\theta,\tau}$, where $F_{t,\mu,s,\theta,\tau}$ is the Pareto change point distribution

$$F_{t,\mu,s,\theta,\tau}(x) = \begin{cases} F_t(x) & \text{if } x \in [x_0, s], \\ 1-(1-F_t(s))(1-G_{s,\mu}(x)) & \text{if } x \in (s,\tau], \\ 1-(1-F_t(s))(1-G_{s,\mu}(\tau))(1-G_{\tau,\theta}(x)) & \text{if } x \in (\tau,\infty), \end{cases} \tag{I.28}$$

where $\mu,\theta > 0$ and $x_0 \le s \le \tau$. For the sake of brevity we denote $\widehat{n}_k = \widehat{n}_{t,h,Y_k}$, where $\widehat{n}_{t,h,\tau}$ is defined in (II.6) . We proceed by consecutive testing for the null hypothesis $\mathcal{H}_0(Y_k)$ against the alternatives $\mathcal{H}_1(Y_k, Y_l)$, for all $k \in [k_0, M_{t,h}]$ and l such that $\delta'\widehat{n}_k \le \widehat{n}_l \le (1-\delta'')\widehat{n}_k$, where $k_0 = \delta_0 M_{t,h} \ge 3$ is a constant interpreted as the initial value of k, and δ_0, δ', δ'' are constants satisfying $0 < \delta_0, \delta', \delta'' < \frac{1}{2}$. The break time $\widehat{s}$ is the first time Y_k for which $\mathcal{H}_0(Y_k)$ is rejected. Recall that $\widehat{\theta}_{t,h,\tau}$ is the maximum likelihood estimator of θ given by (I.12). In the same way we obtain the maximum likelihood estimator of μ :

$$\widehat{\mu}_{t,h,s,t} = \frac{\widehat{n}_{t,h,s}}{\widehat{n}_{t,h,s,\tau}}\widehat{\theta}_{t,h,s} - \frac{\widehat{n}_{t,h,\tau}}{\widehat{n}_{t,h,s,\tau}}\widehat{\theta}_{t,h,\tau},$$

where $\widehat{n}_{t,h,s,\tau} = \sum_{i=1}^{n} W_{t,h}(t_i)\, 1_{\left\{s < X_{t_i} \le \tau\right\}}$. The log-likelihood ratio test statistic for testing $\mathcal{H}_0(s)$ against $\mathcal{H}_1(s,\tau)$ is given by

$$LR_{t,h}(s,\tau) = \widehat{n}_{t,h,s,\tau}\mathcal{K}\left(\widehat{\mu}_{t,h,s,\tau},\widehat{\theta}_{t,h,s}\right) + \widehat{n}_{t,h,\tau}\mathcal{K}\left(\widehat{\theta}_{t,h,\tau},\widehat{\theta}_{t,h,s}\right). \tag{I.29}$$

Taking into account (I.27), we have

$$\mathcal{L}_{t,h}\left(\tau,\widehat{\theta}_{t,h,\tau}\right)-\mathrm{Pen}_{t,h}\left(\tau,\theta\right)=\widehat{n}_{t,h,\tau}\mathcal{K}\left(\widehat{\theta}_{t,h,\tau},\theta\right),$$

which implies that the second term in (I.29) can be viewed as the penalized quasi-log-likelihood

$$\mathcal{L}_{t,h}^{\mathrm{Pen}}\left(s,\tau\right)=\mathcal{L}_{t,h}\left(\tau,\widehat{\theta}_{t,h,\tau}\right)-\mathrm{Pen}_{t,h}\left(\tau,\widehat{\theta}_{t,h,s}\right).$$

We denote by $D>0$ the critical value in the testing procedure below. To speed up the calculations, we take $k=k_0+i\,k_{step}$, $i=0,...,M_{grid}$, where $k_{step}=[M_{t,h}/M_{grid}]$ is an increment for k and M_{grid} is fixed. The values δ_0,k_{step}, δ',δ'' and D are the parameters of the procedure to be determined empirically.

The procedure of the adaptive choice of τ is as follows :

Step 1. Set $k=k_0$.

Step 2. Compute the test statistic

$$\mathbf{Z}_h\left(Y_k\right)=\max_{\delta'\widehat{n}_k\leq\widehat{n}_l\leq(1-\delta'')\widehat{n}_k}LR_{t,h}\left(Y_k,Y_l\right).$$

Step 3. If $k\leq M_{t,h}$ and $\mathbf{Z}_h\left(Y_k\right)>D$, define

$$\widehat{l}=\underset{\delta'\widehat{n}_{k-k_{step}}\leq\widehat{n}_l\leq(1-\delta'')\widehat{n}_{k-k_{step}}}{\arg\max}\mathcal{L}_{t,h}^{\mathrm{Pen}}\left(Y_{\widehat{n}_{k-k_{step}}},Y_l\right) \tag{I.30}$$

and $\widehat{\tau}_n=Y_{\widehat{l}}$. We set the adaptive estimator to $\widehat{\theta}_{t,h,\widehat{\tau}_n}$ and exit the algorithm. If $k\leq M_{t,h}$ and $\mathbf{Z}_h\left(Y_k\right)\leq D$, then : if $k<M_{t,h}-k_{step}$ increase k by k_{step} and return to Step 2. If $k\geq M_{t,h}-k_{step}$, we define $\widehat{l}=M_{t,h}$. Let $\widehat{\tau}_n=Y_{\widehat{l}}$, set the adaptive estimator to $\widehat{\theta}_{t,h,\widehat{\tau}_n}$ and exit the algorithm.

2.4.2 Propagation property of the test statistic

We shall prove that if $Y_k \geq \tau_n$ then the test statistic

$$\mathbf{Z}_h(Y_k) = \sup_{\delta' \widehat{n}_k \leq \widehat{n}_l \leq (1-\delta'')\widehat{n}_k} LR_{t,h}(Y_k, Y_l)$$

does not exceed

$$D = D(n) = c^* \log n \tag{I.31}$$

with high probability for some constant $c^* > 0$.

Theorem I.7. *Assume that the family $(F_t)_{t\in[0,T_{\max}]}$ satisfies condition C1. Then, there exists a constant $c^* > 0$ in (I.31), such that*

$$\mathbf{P}\left(\sup_{Y_k \geq \tau_n} \mathbf{Z}(Y_k) > D(n)\right) \leq \frac{4}{n} \text{ as } n \to \infty,$$

where $\mathbf{Z}(Y_k) = \mathbf{Z}_{h_n}(Y_k)$ and h_n is from condition C1.

Proof. See Section 2.8.4. □

Since $P\left(\sup_{Y_k\geq\tau_n} \mathbf{Z}(Y_k) \leq D(n)\right) \leq P(\widehat{\tau}_n < \tau_n)$, from Theorem I.7 it follows that under condition C1,

$$\mathbf{P}(\widehat{\tau}_n < \tau_n) \to 1 \text{ as } n \to \infty.$$

The meaning of this assertion is that the oracle threshold τ_n is detected by our selection procedure with high probability as $n \to \infty$.

Since condition C1 means that F_t has a Pareto like tail on $[\tau_n, \infty)$, our selection procedure is equivalent to performing a goodness-of-fit test for testing the null hypothesis $\mathcal{H}_0(\widehat{\tau}_n) : F_t = F_{t,\widehat{\tau}_n,\theta}$.

2.4.3 Rates of convergence of the adaptive estimator

We first compare the performance of the adaptive estimator $\widehat{\theta}_{t,h_n,\widehat{\tau}_n}$ with that of the non adaptive estimator $\widehat{\theta}_{t,h_n,\tau_n}$.

Theorem I.8. *Assume that the family* $(F_t)_{t\in[0,T_{\max}]}$ *satisfies condition C1. Then, there exists a constant* $c^* > 0$ *in (I.31), such that as* $n \to \infty$

$$\mathcal{K}(\widehat{\theta}_{t,h_n,\widehat{\tau}_n}, \widehat{\theta}_{t,h_n,\tau_n}) = O_{\mathbf{P}}\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right).$$

Proof. See Section 2.8.5. □

The previous theorem allows to extend the results of the non adaptive setting to the adaptive one. The following theorem gives the rate of convergence of the adaptive estimator $\widehat{\theta}_{t,h_n,\widehat{\tau}_n}$.

Theorem I.9. *Assume that the family* $(F_t)_{t\in[0,T_{\max}]}$ *satisfies condition C1. Then, there exists a constant* $c^* > 0$ *in (I.31), such that as* $n \to \infty$

$$\mathcal{K}(\widehat{\theta}_{t,h_n,\widehat{\tau}_n}, \theta_n) = O_{\mathbf{P}}\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right).$$

Proof. Combining Theorem I.8 and Theorem I.1 we obtain Theorem I.9. □

Recall that the adaptive estimator of the excess distribution function F_{t,τ_n} is given by $G_{\tau_n,\widehat{\theta}_{t,h_n,\widehat{\tau}_n}}$ (see I.10). We give now the rate of convergence of the adaptive estimator $G_{\tau_n,\widehat{\theta}_{t,h_n,\widehat{\tau}_n}}$ to F_{t,τ_n} in terms of the Kullback-Leibler divergence.

Theorem I.10. *Assume that the family* $(F_t)_{t\in[0,T_{\max}]}$ *satisfies condition C1 with* $\theta_n = \theta_{t,\tau_n}$. *Moreover, assume that as* $n \to \infty$

$$\chi^2(F_{t,\tau_n}, G_{\tau_n,\theta_n}) = O\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right). \tag{I.32}$$

Then, there exists a constant $c^* > 0$ *in (I.31), such that as* $n \to \infty$

$$\mathcal{K}\left(F_{t,\tau_n}, G_{\tau_n,\widehat{\theta}_{t,h_n,\widehat{\tau}_n}}\right) = O_{\mathbf{P}}\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right).$$

Proof. See Section 2.8.6. □

In the particular case of the Hall model we obtain an explicit rate of convergence of the adaptive estimator $\widehat{\theta}_{t,h_n,\widehat{\tau}_n}$ (cf. Theorem I.4) as follows :

Theorem I.11. *Under the assumptions of Proposition I.3, there exists a constant $c^* > 0$ in (I.31), such that*

$$\sqrt{\mathcal{K}\left(\widehat{\theta}_{t,h_n,\widehat{\tau}_n}, \theta_{t,\tau_n}\right)} = O_{\mathbf{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{1+\beta(2+1/\rho)}}\right) \quad as \quad n \to \infty,$$

where $h_n \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{1+\beta(2+1/\rho)}}$.

Proof. This Theorem is a consequence of Theorem I.8 and Theorem I.4. □

In the case of mixture of two Pareto distributions, we obtain similar rate of convergence for the adaptive estimator.

2.5 Simulations

We first give arguments on the choice of the parameters of the selection procedure given in Section 2.4. The proposed procedure depends on the initial proportion δ_0, the parameters δ', δ'', the critical value D and the grid length $M_{grid} = m/k_{step}$, with m being the number of observations in the window $[t-h_n, t+h_n]$. The parameters δ_0, δ' and δ'' should be large enough to prevent from large variability in the first several iterations of the algorithm. We fix $\delta_0 = \frac{1}{10}$, $\delta' = \frac{1}{4}$ and $\delta'' = \frac{1}{20}$. Our simulations show that the procedure is not very sensitive to the choice of the parameter M_{grid}. We choose $M_{grid} = 100$ to reduce the computation time. We choose as kernel function the truncated gaussian density function given by

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) 1_{[-1,1]}(x), \tag{I.33}$$

but we have also performed computations for other kernels like Gaussian, Epanechnikov and uniform.

Under the null hypothesis the distribution function of the test statistic

$$T_n = \sup_{k_0 \le k = k_0 + i k_{step} \le n} \mathbf{Z}_{h_n}(Y_k)$$

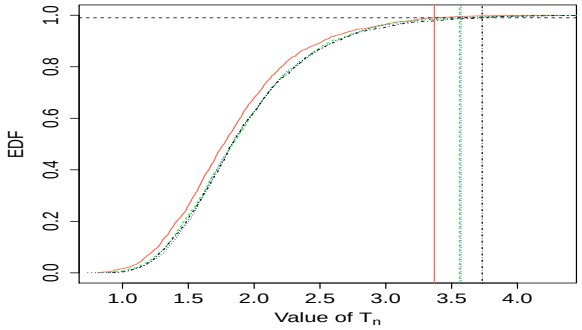

Figure I.1 – Empirical Distribution Function (EDF) of the statistic T_n when the number of observations m in the window $[t-h_n, t+h_n]$ is varying from 200 to 2000. Red line corresponds to $m = 200$, blue line to $m = 500$, green line to $m = 1000$ and black line to $m = 2000$. The dashed horizontal black line represents the 0.99-probability value and the vertical dashed lines correspond to the respective critical values.

does not depend on the unknown parameter θ. Therefore, to compute the critical value D of T_n, we can fix the value of θ equal to 1. To determine D, we simulate the values of the test statistic T_n under the null hypothesis. The value D is chosen as the 0.99-empirical quantile to ensure a 0.01 type I error. This choice is motivated by the propagation property of the test statistic under the null hypothesis which guaranties that T_n does not exceed the critical value D with high probability.

The empirical distribution functions of T_n for the choice of kernel (I.33) and various sample sizes in the window $[t-h_n, t+h_n]$ is given in Figure I.1. These numerical results show that the empirical distribution function of T_n does not depend much on the number of observations m in the window $[t-h_n, t+h_n]$ and therefore the corresponding 0.99-level critical values $D = D_{0.99}$ are nearly the same. Our simulations also show that the threshold $\tau_n = Y_{\widehat{l}}$ in (II.35) is not very sensitive to the choice of the critical value D in the sense that the adaptive choice $\widehat{l}$ remains constant with respect to relatively large variations of D. The chosen value for D is fixed to 3.6.

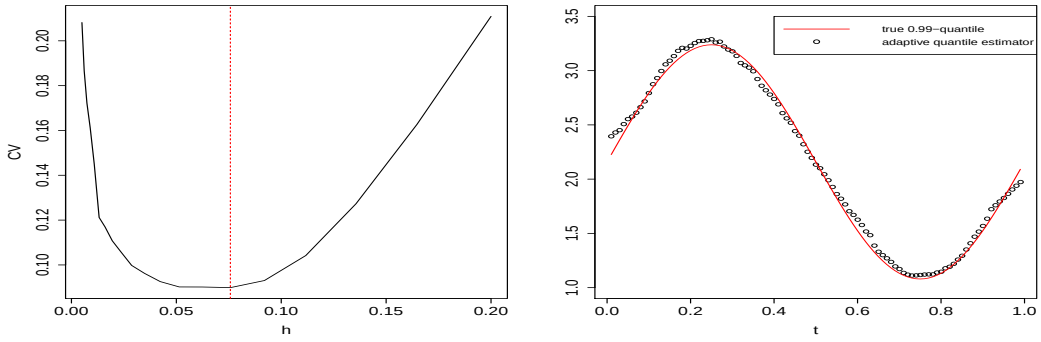

Figure I.2 – Left : Representation of the cross validation function $CV(h, p = 0.99)$. The dashed vertical line corresponds to the optimal value $h_{cv} = 0.76$. Right : Representation of the log of the true 0.99-quantile (red line) and of the log of the adaptive quantile estimator (black points) in the vertical axis.

To study the behavior of the adaptive estimator under the alternative hypothesis, we consider the mixture model (see Section 2.3.3)

$$F_t(x) = C(1 - x^{-1/\theta_t}) + (1 - C)\left(1 - x^{-1/\theta_t - 5}\right), \; x \geq 1, 0 \leq t \leq 1 \tag{I.34}$$

where $C = 0.75$ and

$$\theta_t = 0.5 + 0.25 \sin(2\pi t).$$

We take $n = 50000$ to keep nearly the same number of observations as in the application given in Section 2.6. We generate $N_{MC} = 1000$ replicates of $X_{t_1}, ..., X_{t_n}$ with $t_i = \frac{i}{n}, i = 1, ..., n$ from the mixture model (I.34). We focus on the estimation of $F_{t_i}^{-1}(p)$ for t_i in the uniform grid $T_{grid} = \{0.01, 0.02, \ldots, 0.99\}$. The estimators $\widehat{\theta}_{t_i, h_n, \widehat{\tau}_n}$ are computed using the adaptive procedure from Section 2.4.1 with the parameters δ_0, δ', δ'', M_{grid} and D fixed above.

We fix the probability level $p = 0.99$. To select the bandwidth h_n we consider the sequence $h_l = aq^l$, $l = 0, \ldots, M_h = 20$ with $a = 0.005$, $b = 0.2$ and $q = \exp\left(\frac{\log b - \log a}{M_h}\right)$. We choose

$h_n = h_{cv}$ by minimizing in h_m, $m = 1, \ldots, M_h$ the cross validation function

$$CV(h_m, p) = \frac{1}{M_h card(T_{grid})} \sum_{h_l} \sum_{t_i \in T_{grid}} \psi\left(\widehat{F}^{-1}_{t_i,h_l}(p), \widehat{q}^{(-i)}_p(t_i, h_m)\right), \tag{I.35}$$

where $\psi(x, y) = |\log x - \log y|$ is the least absolute relative deviation loss function and the quantile estimator $\widehat{q}^{(-i)}_p(\cdot)$ is defined by (I.15) with the observation X_{t_i} removed.

In Figure I.2 (right picture), we plot one realization of the adaptive quantile estimator with the bandwidth h_{cv} selected by cross validation function (I.35) (left picture). We note that the approximation provided by our adaptive estimator is very close to the true 0.99-quantile. This is also confirmed by Figure I.3, where we give the boxplots of the log of the values $\widehat{q}_p(t_i)$ (left picture) and the empirical Mean Squared Relative Error (MSRE) (right picture) from $N_{MC} = 1000$ realizations. Here the MSRE is given by

$$\text{MSRE}(t, p) = \frac{1}{N_{MC}} \sum_{m=1}^{N_{MC}} \log^2\left(\widehat{q}^{(m)}_p(t, h^{(m)}_{cv})/F^{-1}_t(p)\right)$$

where $h^{(m)}_{cv}$ and $\widehat{q}^{(m)}_p(t, h^{(m)}_{cv})$ are respectively the cross validation bandwidth and the adaptive quantile estimator on m-th Monte Carlo simulation. As expected, the heavier is the tail, the larger is the MSRE : the MSRE on the interval $[0.1, 0.4]$ is larger than on the interval $[0.5, 0.9]$. At the ends of the interval we observe the largest errors due to the boundary effect.

2.6 Application : monitoring aquatic biosensors

Protection of the aquatic environment is a top priority for marine managers, policy makers, and the general public. Human activities are responsible for significant discharges of pollutants into environment. These pollutants lead to the degradation of many habitats disturbing ecosystems and also causing problems in terms of public health. Surveillance and protection of aquatic systems is thus fundamental and it is of great interest to be able to inform in real time people of water conditions. Due to an increasing interest in the health of aquatic systems,

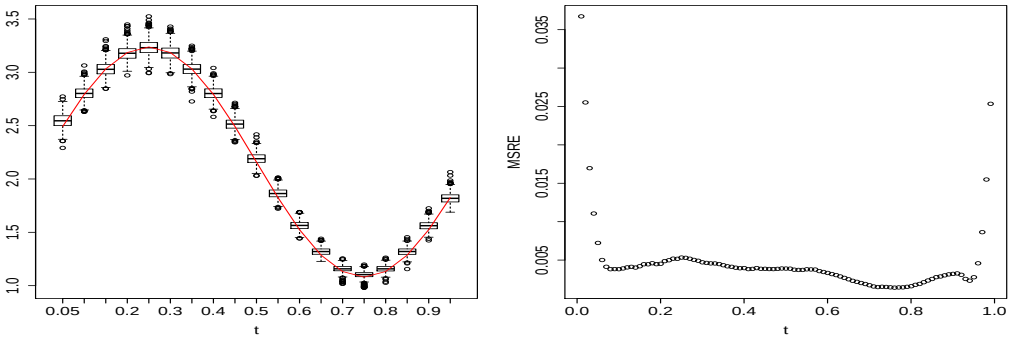

Figure I.3 – Boxplots (left picture) and MSRE (right picture) of the log of the adaptive estimator $\widehat{q}_p(t)$ for $t \in [0,1]$ for the Hall model from $N_{MC} = 1000$ realizations and $p = 0.99$. In the left picture, the red line represents the log of the true 0.99-quantile.

there is a compelling need for the use of remote online sensors to instantly and widely distribute information on a daily basis. Regulations and controls on water quality have already been established. Among these controls, bioindicators are increasingly used because they can be effective in their ability to reveal the presence of traces (very low concentrations) of contaminants through accumulation in tissues of aquatic animals (see [4, 5, 49]).

The interest in investigating the bivalve's activities by recording their valve movements (valvometry) has been explored in ecotoxicology. The basic idea of valvometry is to use the bivalve's ability to close its shells when exposed to a contaminant as an alarm signal ([8, 50, 51] among others). There has been a clear research interest in the recent years to measuring the bivalve's behaviors directly in real conditions ([4, 8, 52]).

These noninvasive valvometric techniques provide high-frequency data and different statistical models were proposed to study their behavior in their natural habitat and constantly monitor water quality when faced with stress such as a pollutant : valves can suddenly close or express abnormal movements indicating a change in water quality ([8–11, 53]). When faced with pollution or poor quality water, the oyster closes its shells, or in extreme situations the

activity of animals can change dramatically after exposure to even very low levels of pollution.

We first describe the experimental site and the animal species. Then, we give some details on evaluation of valve activity. Afterwards, we provide information on data collection and transmission. Finally, exploring typical features of the valvometric environmental data samples (i.e. measurements of distances between the two parts of the shell of bivalves) collected by a laboratory called Environnements et Paléoenvironnements Océaniques et Continentaux (EPOC, http://www.epoc.u-bordeaux.fr/), we explain which inferences are valuable from the biological point of view.

2.6.1 Data acquisition

The monitoring site we considered is located in France at Locmariaquer (Latitude : 47°57 N, Longitude : 2°94 W). A group of sixteen Pacific oysters, *Crassostrea gigas*, measuring from 8 to 10 cm length, are installed on each site. Every oyster has almost the same age (1.5 years old). They were placed in a traditional oyster farmer bag.

The electronic principle of monitoring was described by [4] and further modified by [6]. Some information about these specific aspects can be found on http://molluscan-eye.epoc.u-bordeaux1.fr. In brief, each animal is equipped with two light coils (sensors), of approximately 53 mg each (unembedded), fixed on the edge of each valve. These coils measure 2.5×2.5×2 mm and were coated with a resin sealing before fixation on the valves. One of the coils emits a high-frequency, sinusoidal signal which is received by the other coil. For each sixteen animals, one measurement is received every 0.1s (10 Hz). This means that each animal's behavior is measured every 1.6s. Every day, a data set with 864,000 pairs of values (1 distance value, 1 stamped time value) is generated. A first electronic card manages the electrodes and is in a waterproof case next to the animals. A second electronic card handling the data acquisition and the programmed emission is also in the field but outside the water on a pier. This unit

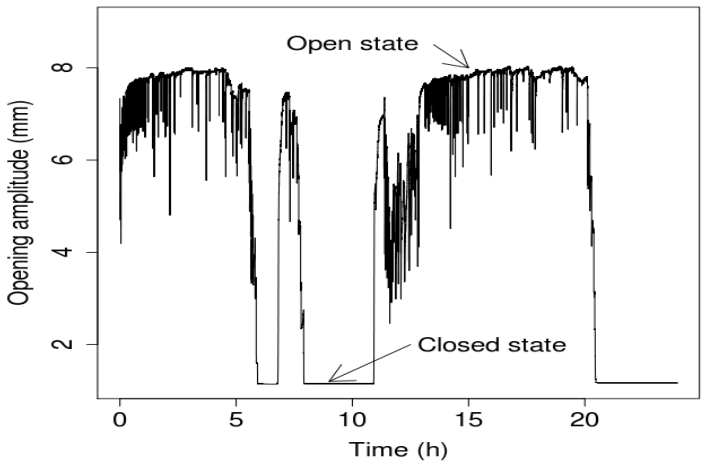

Figure I.4 – An example of valvometric data for one oyster

is equipped with a GSM/GPRS modem and uses Linux operating system for driving the first control unit, managing the data storage, accessing the Internet, and transferring the data.

A self-developed software module runs on mobile phone technology. After each 24h period (or any other programmed period of time), the data collected are transmitted to a remote central workstation server where they are stored in text files. Every day, files from each site are inserted in a SQL database which is accessible with the software R [54] or a text terminal, via Internet or directly from the storage server.

These measurements produce some characteristic features that can be examined in Figure I.4. As argued in [4, 8, 10, 55], pollution can affect the activity of oysters and in particular the shells opening and closing speed. For instance in an inhospitable extreme environment, oysters will close more rapidly its shells. Thus, detecting extreme changes of the closing speed can provide insights about the health of oysters and so give an insight about the water quality.

2.6.2 Application results

We consider dataset associated to movement speeds which are considered as an indicator of the animal stress activity since its movements are associated to aquatic system perturbations.

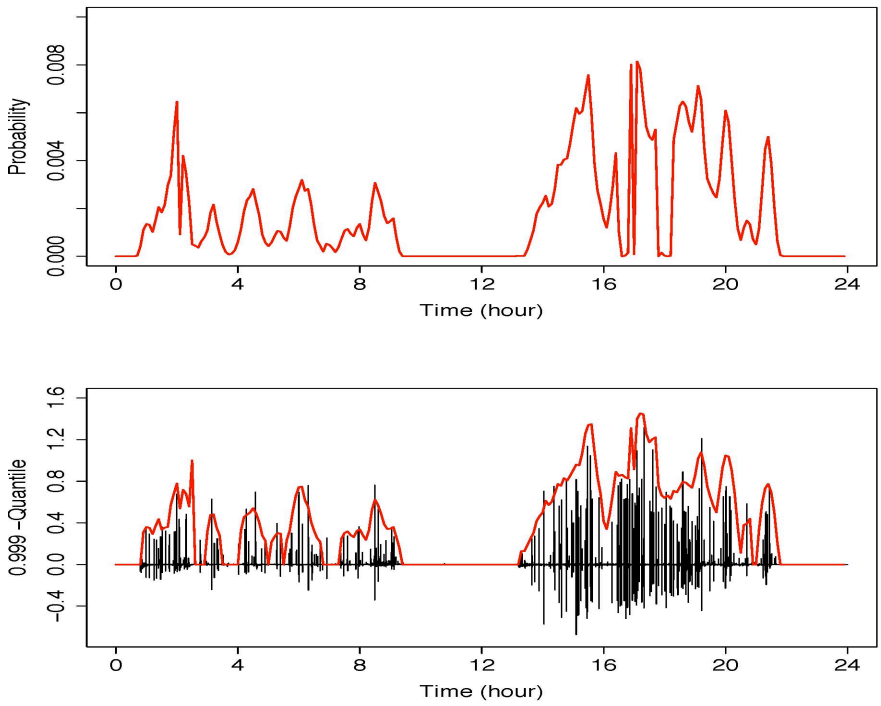

Figure I.5 – The red line on the top figure displays the estimated tail probabilities $P(X_t > 0.3)$ on April 18, 2011. The red line on the bottom figure displays the 0.999-quantile process the same day. The black lines represent the speeds of valve closings.

For instance, a stressed oyster in the presence of pollution or environmental perturbations exhibits irregular and numerous micro-closings and opening periods with high speed.

Figure IV.4 shows for the 18th of April 2011 the plot of probability $P(X_t > 0.3)$ and 0.999-quantile estimators of the valve closing speed for one oyster in the Locmariaquer site. For an easier data visualization of the extreme quantiles of the closing velocities of the 16 oysters through the period starting from 4th of March to 21th of August 2011, we use in Figure I.6 a customized color table (gray color associated to the smallest quantiles class, yellow to the intermediate quantiles class and red to the largest quantiles class) to match computed extreme quantile values.

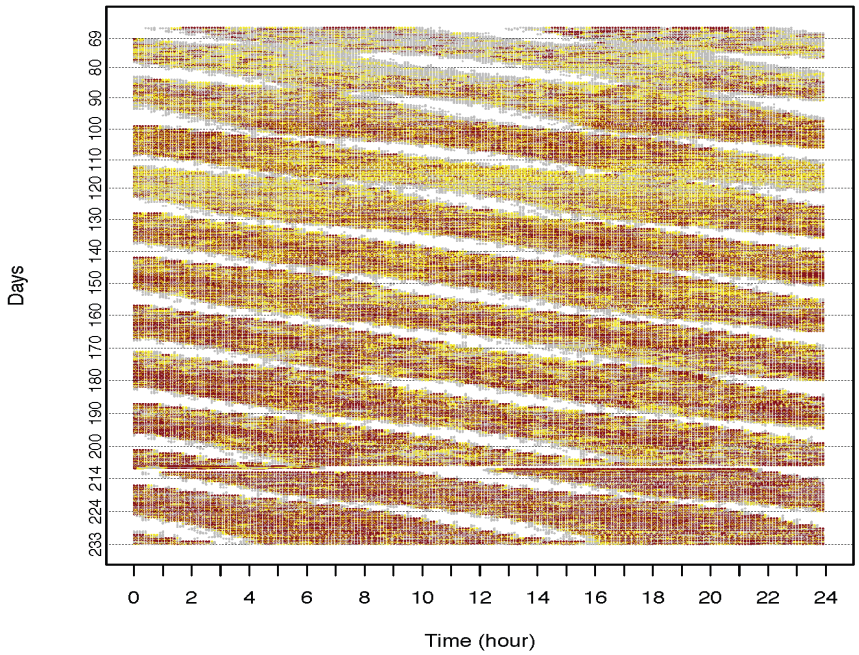

Figure I.6 – Representation of the extreme 0.999-quantile estimator of the closing velocity between the 4th of March and the 21th of August 2011 considering the 16 oysters in Locmariaquer. The x-axis represents the time in a 24 hour time period and the y-axis represents the number of days since the 1st of January 2011

.

The 4th of March and the 21th of August 2011 correspond respectively to the 63th days and to the 233th days of the 2011 year. For each day, there are 16 lines of colored points representing the extreme 0.999-quantiles values at each $t \in [0, 24]$ hour of each oyster's velocity. The advantage of this representation is to give the extreme quantiles values for each of the 16 oysters for a given time period in one single graph. Figure I.6 shows that the closing activity is highly correlated with the tidal amplitude and that the closing state is synchronized with the low tide period. This is confirmed by [8, 10] using non parametric methods.

We notice particularly a yellow zone (between the 110th and 125th days) explained by a

sudden change in temperature collected by a temperature sensor installed near the oysters (data not shown) and a red colored area showing a more intense activity including spawning activity (days ≥ 210). Thus these results contribute to the development of a tool for monitoring water quality based on the analysis of continuous behavior of bivalves (bio-indicator of pollution in the field). This velocity information provides an important indication of the change in behavior of oysters such as a spawn or a period of abnormal stress characterized by rapid partial closures and openings).

2.7 Conclusion

Theoretical results : We prove the convergence of the estimators of the adjusted model parameters when the threshold and the kernel bandwidth are deterministic and we give their rates of convergence. These convergence results are then extended to the corresponding estimators with adaptive threshold.

Adaptive choice of parameters : In applications the threshold and the bandwidth are usually not known. We give a selection procedure based on maximal propagation of the parametric adjustment and a subsequent choice of the threshold parameter using penalized maximum likelihood. We propose a cross validation method to determine the bandwidth parameter.

Model validity : The construction of the adaptive estimator is based on a testing procedure which can be viewed as a goodness-of-fit test for the parametric-based part of the model. So the question of the model validation for the adjusted tail is answered by our adaptive procedure : at each step of the procedure the adjusted tail is tested and if it is not rejected the sample is enlarged and tested again until the parametric model is rejected. The choice of the "optimal" threshold is made among the already tested models and therefore the adaptive adjusted tail is validated as well. If the test rejects the parametric tail fit from the very beginning, the Pareto tail adjustment is not significant. On the opposite, if all the tests accept the parametric Pareto

fit then the underlying distribution is significantly Pareto.

Simulations : We study the behavior of the estimators under the null and the alternative hypotheses. Under the null hypothesis, assuming i.i.d. standard Pareto observations, we compute the critical value in the adaptive procedure and we show that it remains stable with respect to the number of observations in the window. This is done for the truncated Gaussian kernel, but the critical values for other kernels can be determined in the same way. We perform numerical simulations with the adaptive estimators of the Pareto tail parameter in order to show that under the alternative hypothesis the relative mean squared error is small.

Application : We apply the developed procedure in the context of an ecological study. The objective is to determine extreme environmental disturbances through high frequency measurements of oysters activity considered as a bioindicator of pollution.

2.8 Proofs

We consider a semiparametric Pareto model with two change points v and τ defined by

$$F_{t,\theta,s,\mu,v,\theta',\tau}(x) = \begin{cases} F_t(x) & \text{if } x \in [x_0, s], \\ 1-(1-F_t(s))(1-G_{s,\theta}(x)) & \text{if } x \in (s, v], \\ 1-(1-F_t(s))(1-G_{s,\theta}(v))(1-G_{v,\mu}(x)) & \text{if } x \in (v, \tau], \\ 1-(1-F_t(s))(1-G_{s,\theta}(v))(1-G_{v,\mu}(\tau))(1-G_{\tau,\theta'}(x)) & \text{if } x > \tau, \end{cases}$$

where $t \in [0,1]$, $\theta', \theta, \mu > 0$ and $\tau \geq v \geq s \geq x_0$. In the following we consider a bandwidth parameter $h > 0$. Let

$$\begin{aligned} \mathcal{L}_{t,h}(F_{t,v,\theta}) &= \mathcal{L}_{t,h}(v,\theta) = \sum_{i=1}^{n} W_{t,h}(t_i) \log \frac{dF_{t,v,\theta}}{dx}(X_{t_i}), \\ \mathcal{L}_{t,h}(F_{t,\mu,v,\theta,\tau}) &= \sum_{i=1}^{n} W_{t,h}(t_i) \log \frac{dF_{t,\mu,v,\theta,\tau}}{dx}(X_{t_i}), \\ Z_{t,h}(F_{t,v,\theta'}, F_{t,v,\theta}) &= \mathcal{L}_{t,h}(F_{t,v,\theta'}) - \mathcal{L}_{t,h}(F_{t,v,\theta}) = \sum_{i=1}^{n} W_{t,h}(t_i) \log \frac{dF_{t,v,\theta'}}{dF_{t,v,\theta}}(X_{t_i}), \end{aligned}$$

and

$$Z_{t,h}(F_{t,\mu,v,\theta,\tau}, F_{t,v,\theta}) = \mathcal{L}_{t,h}(F_{t,\mu,v,\theta,\tau}) - \mathcal{L}_{t,h}(F_{t,v,\theta}) = \sum_{i=1}^{n} W_{t,h}(t_i) \log \frac{dF_{t,\mu,v,\theta,\tau}}{dF_{t,v,\theta}}(X_{t_i}).$$

The following proposition gives the exponential bounds for the quasi-log-likelihood ratios $Z_{t,h}(F_{t,v,\theta'}, F_{t,v,\theta})$ and $Z_{t,h}(F_{t,\mu,v,\theta,\tau}, F_{t,v,\theta})$. We introduce a measure of discrepancy between the family of distributions $(F_u)_{u\in[0,T_{\max}]}$ and the adjusted models $(F_{u,s,\theta})_{u\in[0,T_{\max}]}$ at time t by

$$d_{t,h,s,\theta} = \sum_{i=1}^{n} W_{t,h}(t_i)\chi^2(F_{t_i}, F_{t_i,s,\theta}).$$

Proposition I.12. *For any $y > 0, \tau \geq v \geq s \geq x_0$, and any $\mu, \theta, \theta' > 0$, we have*

$$\mathbf{P}\left(Z_{t,h}(F_{t,v,\theta'}, F_{t,v,\theta}) > y\right) \leq \exp\left(-\frac{y}{2} + \frac{d_{t,h,s,\theta}}{2}\right), \tag{I.36}$$

$$\mathbf{P}\left(Z_{t,h}(F_{t,\mu,v,\theta,\tau}, F_{t,v,\theta}) > y\right) \leq \exp\left(-\frac{y}{2} + \frac{d_{t,h,s,\theta}}{2}\right). \tag{I.37}$$

Proof. We first prove (I.36). Let $v \geq s \geq x_0$. Since $\frac{dF_{t,v,\theta'}}{dF_{t,v,\theta}} = \frac{dF_{t_i,\theta,s,\theta',v}}{dF_{t_i,s,\theta}}$ for $i = 1, ..., n$, by the definition of $Z_{t,h}(F_{t,v,\theta'}, F_{t,v,\theta})$, we have

$$Z_{t,h}(F_{t,v,\theta'}, F_{t,v,\theta}) = \sum_{i=1}^{n} W_{t,h}(t_i) \log \frac{dF_{t_i,\theta,s,\theta',v}}{dF_{t_i,s,\theta}}(X_{t_i}).$$

Denote for brevity $H_{t_i} = F_{t_i,\theta,s,\theta',v}$. Applying exponential Chebyshev's inequality, we have

$$\mathbf{P}\left(Z_{t,h}(F_{t,v,\theta'}, F_{t,v,\theta}) > y\right) \leq \exp\left(-\frac{y}{2}\right) \mathbf{E}\left(\exp\left(\frac{1}{2} Z_{t,h}(F_{t,v,\theta'}, F_{t,v,\theta})\right)\right). \tag{I.38}$$

Since $0 \leq W_{t,h}(t_i) \leq 1$ for all $i = 1, ..., n$, we deduce by Hölder's inequality

$$\begin{aligned}
&\log \mathbf{E}\left(\exp\left(\frac{1}{2} Z_{t,h}(F_{t,v,\theta'}, F_{t,v,\theta})\right)\right) \\
&= \sum_{i=1}^{n} \log \mathbf{E}\left(\exp\left(\frac{1}{2} 1_{\{X_{t_i} > s\}} \log \frac{dH_{t_i}}{dF_{t_i,s,\theta}}(X_{t_i})\right)\right)^{W_{t,h}(t_i)} \\
&\leq \sum_{i=1}^{n} W_{t,h}(t_i) \log \mathbf{E}\left(\exp\left(\frac{1}{2} 1_{\{X_{t_i} > \tau\}} \log \frac{dH_{t_i}}{dF_{t_i,s,\theta}}(X_{t_i})\right)\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{E}\left(\exp\left(\frac{1}{2}1_{\{X_{t_i}>s\}}\log\frac{dH_{t_i}}{dF_{t_i,s,\theta}}(X_{t_i})\right)\right)\\
=\ &\mathbf{E}\left(\exp\left(\frac{1}{2}1_{\{X_{t_i}>s\}}\log\frac{dH_{t_i}}{dF_{t_i}}(X_{t_i})\right)\exp\left(\frac{1}{2}1_{\{X_{t_i}>s\}}\log\frac{dF_{t_i}}{dF_{t_i,s,\theta}}(X_{t_i})\right)\right)\\
\leq\ &\sqrt{\mathbf{E}\left(\exp\left(1_{\{X_{t_i}>s\}}\log\frac{dH_{t_i}}{dF_{t_i}}(X_{t_i})\right)\right)}\sqrt{\mathbf{E}\left(\exp\left(1_{\{X_{t_i}>s\}}\log\frac{dF_{t_i}}{dF_{t_i,s,\theta}}(X_{t_i})\right)\right)}.
\end{aligned}$$

Using the fact that, for $i=1,...,n$

$$\mathbf{E}\left(\exp\left(1_{\{X_{t_i}>s\}}\log\frac{dH_{t_i}}{dF_{t_i}}(X_{t_i})\right)\right)=1$$

and

$$\mathbf{E}\left(\exp\left(1_{\{X_{t_i}>s\}}\log\frac{dF_{t_i}}{dF_{t_i,s,\theta}}(X_{t_i})\right)\right)=1+\chi^2(F_{t_i},F_{t_i,s,\theta}),$$

we obtain

$$\begin{aligned}
\log\left(\mathbf{E}\left(\exp\left(\frac{1}{2}Z_{t,h}(F_{t,v,\theta'},F_{t,v,\theta})\right)\right)\right) &\leq \frac{1}{2}\sum_{i=1}^{n}W_{t,h}(t_i)\log\left(1+\chi^2(F_{t_i},F_{t_i,s,\theta})\right)\\
&\leq \frac{1}{2}\sum_{i=1}^{n}W_{t,h}(t_i)\chi^2(F_{t_i},F_{t_i,s,\theta})=\frac{d_{t,h,s,\theta}}{2}. \qquad \text{(I.39)}
\end{aligned}$$

Combining (I.38) and (I.39), we deduce that

$$\mathbf{P}\left(Z_{t,h}(F_{t,v,\theta'},F_{t,v,\theta})>y\right)\leq\exp\left(-\frac{y}{2}+\frac{d_{t,h,s,\theta}}{2}\right), \qquad \text{(I.40)}$$

which prove that (I.36) is satisfied.

Since, for $i=1,...,n$

$$\frac{dF_{t,\mu,v,\theta,\tau}}{dF_{t,v,\theta}}(X_{t_i})=\frac{dF_{t_i,\theta,s,\mu,v,\theta,\tau}}{dF_{t_i,s,\theta}}(X_{t_i}),$$

we have

$$Z_{t,h}(F_{t,\mu,v,\theta,\tau},F_{t,v,\theta})=\sum_{i=1}^{n}W_{t,h}(t_i)\log\frac{dF_{t_i,\theta,s,\mu,v,\theta,\tau}}{dF_{t_i,s,\theta}}(X_{t_i}).$$

Now (I.37) is proved in the same way as (I.36). □

Next, we give an exponential bound for the maximum quasi-log-likelihood ratio which permits to obtain a rate of convergence of nonadaptive estimator $\widehat{\theta}_{t,h,\tau_n}$.

Proposition I.13. *For any $y > 0, \tau \geq v \geq s \geq x_0$, and $\theta > 0$, we have*

$$\begin{aligned}
\mathbf{P}\left([\widehat{n}_{t,h,v}]\mathcal{K}(\widehat{\theta}_{t,h,v},\theta) > y\right) &\leq 2n\exp\left(-\frac{y}{2}+\frac{d_{t,h,s,\theta}}{2}\right),\\
\mathbf{P}\left([\widehat{n}_{t,h,v,\tau}]\mathcal{K}(\widehat{\mu}_{t,h,v,\tau},\theta) > y\right) &\leq 2n\exp\left(-\frac{y}{2}+\frac{d_{t,h,s,\theta}}{2}\right),
\end{aligned}$$

where $[u]$ is the integer part of u.

Proof. We shall prove only the first inequality, the second one being proved in the same way. We start with the obvious relation $Z_{t,h}(F_{t,v,\theta'}, F_{t,v,\theta}) = \widehat{n}_{t,h,v}\Lambda(\theta')$, where $\Lambda(u) = \log\frac{\theta}{u} - (\frac{1}{u}-\frac{1}{\theta})\widehat{\theta}_{t,h,v}$. Denote for brevity $g(u,k) = \left(\log\frac{\theta}{u} - \frac{y}{k}\right)/\left(\frac{1}{u}-\frac{1}{\theta}\right)$. Note that for $k > 0$ and $0 < u < \theta$, the inequality $k\Lambda(u) > y$ is equivalent to $g(u,k) > \widehat{\theta}_{t,h,v}$ for $k > 0$. Similarly, for $u > \theta$ the inequality $k\Lambda(u) > y$ is equivalent to $g(u,k) < \widehat{\theta}_{t,h,v}$. Moreover with $k > 0$ fixed, the function $g(u,k)$ has a maximum for $0 < u < \theta$ and a minimum for $u > \theta$. Let $\theta^+(k) = \arg\max_{0<u<\theta} g(u,k)$ and $\theta^-(k) = \arg\min_{u>\theta} g(u,k)$. One can see that

$$\begin{aligned}
\left\{[\widehat{n}_{t,h,v}]\,\Lambda(\widehat{\theta}_{t,h,v}) > y, \widehat{\theta}_{t,h,v} < \theta\right\} &= \left\{g(\widehat{\theta}_{t,h,v},[\widehat{n}_{t,h,v}]) > \widehat{\theta}_{t,h,v}, \widehat{\theta}_{t,h,v} < \theta\right\}\\
&\subseteq \left\{g(\theta^+([\widehat{n}_{t,h,v}]),[\widehat{n}_{t,h,v}]) > \widehat{\theta}_{t,h,v}, \widehat{\theta}_{t,h,v} < \theta\right\}\\
&= \left\{[\widehat{n}_{t,h,v}]\,\Lambda(\theta^+([\widehat{n}_{t,h,v}]) > y, \widehat{\theta}_{t,h,v}, < \theta\right\}\\
&\subseteq \left\{[\widehat{n}_{t,h,v}]\,\Lambda(\theta^+([\widehat{n}_{t,h,v}]) > y\right\}.
\end{aligned}$$

In the same way, we have

$$\left\{[\widehat{n}_{t,h,v}]\,\Lambda(\widehat{\theta}_{t,h,v}) > y, \widehat{\theta}_{t,h,v}, > \theta\right\} \subseteq \left\{[\widehat{n}_{t,h,v}]\,\Lambda(\theta^-([\widehat{n}_{t,h,v}]) > y\right\}.$$

Since $\Lambda(\widehat{\theta}_{t,h,v}) = \mathcal{K}(\widehat{\theta}_{t,h,v},\theta)$ for any $\theta > 0$ and $\mathcal{K}(\widehat{\theta}_{t,h,v},\theta) = 0$ for $\theta = \widehat{\theta}_{t,h,v}$, these inclusions imply

$$\left\{[\widehat{n}_{t,h,v}]\,\mathcal{K}(\widehat{\theta}_{t,h,v},\theta) > y\right\} \subseteq \left\{[\widehat{n}_{t,h,v}]\,\Lambda(\theta^+([\widehat{n}_{t,h,v}]) > y\right\} \cup \left\{[\widehat{n}_{t,h,v}]\,\Lambda(\theta^-([\widehat{n}_{t,h,v}]) > y\right\}.$$

Hence,

$$\begin{aligned}
&\mathbf{P}\left([\widehat{n}_{t,h,v}]\,\mathcal{K}(\widehat{\theta}_{t,h,v},\theta) > y\right)\\
\leq\ &\mathbf{P}\left([\widehat{n}_{t,h,v}]\,\Lambda(\theta^+([\widehat{n}_{t,h,v}]) > y\right) + \mathbf{P}\left([\widehat{n}_{t,h,v}]\,\Lambda(\theta^-([\widehat{n}_{t,h,v}]) > y\right)\\
\leq\ &\sum_{k=1}^{[n_{t,h}]}\mathbf{P}\left([\widehat{n}_{t,h,v}]\,\Lambda(\theta^+(k)) > y\right) + \sum_{k=1}^{[n_{t,h}]}\mathbf{P}\left([\widehat{n}_{t,h,v}]\,\Lambda(\theta^-(k)) > y\right),
\end{aligned}$$

where $n_{t,h} = \sum_{i=1}^{n} W_{t,h}(t_i)$. From Proposition I.12, and the inclusion $\{[\widehat{n}_{t,h,v}]\,\Lambda(\theta') > y\} \subseteq$

$\{\widehat{n}_{t,h,v}\Lambda(\theta') > y\}$, we have, for any $\theta' > 0$

$$\begin{aligned}\mathbf{P}\left([\widehat{n}_{t,h,v}]\,\Lambda(\theta') > y\right) &\leq \mathbf{P}\left(\widehat{n}_{t,h,v}\Lambda(\theta') > y\right)\\ &= \mathbf{P}(Z_{t,h}(F_{t,v,\theta'}, F_{t,v,\theta}) > y),\\ &\leq \exp\left(-\frac{y}{2} + \frac{d_{t,h,s,\theta}}{2}\right).\end{aligned}$$

We deduce

$$\mathbf{P}\left([\widehat{n}_{t,h,v}]\,\mathcal{K}(\widehat{\theta}_{t,h,v}, \theta) > y\right) \leq 2[n_{t,h}]\exp\left(-\frac{y}{2} + \frac{d_{t,h,s,\theta}}{2}\right).$$

Since $[n_{t,h}] \leq n$, we have

$$\mathbf{P}\left([\widehat{n}_{t,h,v}]\,\mathcal{K}(\widehat{\theta}_{t,h,v}, \theta) > y\right) \leq 2n\exp\left(-\frac{y}{2} + \frac{d_{t,h,s,\theta}}{2}\right).$$

□

Proposition I.14. *For any $y > 0$, $s \geq x_0$ and $\theta > 0$, we have*

$$\mathbf{P}\left(\sup_{s\leq v}[\widehat{n}_{t,h,v}]\mathcal{K}(\widehat{\theta}_{t,h,v}, \theta) > y\right) \leq 2n^4\exp\left(-\frac{y}{2} + \frac{d_{t,h,s,\theta}}{2}\right) + \frac{1}{n}$$

and

$$\mathbf{P}\left(\sup_{s\leq v\leq\tau}[\widehat{n}_{t,h,v,\tau}]\mathcal{K}(\widehat{\mu}_{t,h,v,\tau}, \theta) > y\right) \leq n^7\exp\left(-\frac{y}{2} + \frac{d_{t,h,s,\theta}}{2}\right) + \frac{1}{n}.$$

Proof. The proof of the proposition is similar that of Proposition 7.4 in [15]. □

The following Proposition gives an exponential bound for the statistic $LR_{t,h}(v,\tau)$ (see (I.29)

).

Proposition I.15. *For any $y > 0$, $s \geq x_0$ and $\theta > 0$, we have*

$$\mathbf{P}\left(\sup_{s\leq v\leq\tau} LR_{t,h}(v,\tau) > 4y\right) \leq 2n^7\exp\left(-\frac{y}{2} + \frac{d_{t,h,s,\theta}}{2}\right) + \frac{2}{n}.$$

Proof. From the fact that

$$LR_{t,h}(v,\tau) = \max_{\mu,\theta'>0}\mathcal{L}_{t,h}(F_{t,\mu,v,\theta',\tau}) - \max_{\theta>0}\mathcal{L}_{t,h}(F_{t,v,\theta})$$

and

$$\max_{\theta>0}\mathcal{L}_{t,h}(F_{t,v,\theta}) \geq \mathcal{L}_{t,h}(F_{t,v,\theta}),$$

it follows that

$$LR_{t,h}(v,\tau) \leq \max_{\mu,\theta'>0}\mathcal{L}_{t,h}(F_{t,\mu,v,\theta',\tau}) - \mathcal{L}_{t,h}(F_{t,v,\theta}).$$

Proceeding as in the proof of (I.29), we see that

$$\begin{aligned}\max_{\mu,\theta'>0} \mathcal{L}_{t,h}(F_{\mu,v,\theta',\tau}) - \mathcal{L}_{t,h}(F_{t,v,\theta}) &= \widehat{n}_{t,h,v,\tau}\mathcal{K}(\widehat{\mu}_{t,h,v,\tau},\theta) + \widehat{n}_{t,h,\tau}\mathcal{K}(\widehat{\theta}_{t,h,\tau},\theta) \\ &\leq 2[\widehat{n}_{t,h,v,\tau}]\mathcal{K}(\widehat{\mu}_{t,h,v,\tau},\theta) + 2[\widehat{n}_{t,h,\tau}]\mathcal{K}(\widehat{\theta}_{t,h,\tau},\theta).\end{aligned}$$

We deduce

$$LR_{t,h}(v,\tau) \leq 2[\widehat{n}_{t,h,v,\tau}]\mathcal{K}(\widehat{\mu}_{t,h,v,\tau},\theta) + 2[\widehat{n}_{t,h,\tau}]\mathcal{K}(\widehat{\theta}_{t,h,\tau},\theta)$$

and

$$\begin{aligned}\left\{\sup_{s\leq v\leq\tau} LR_{t,h}(v,\tau) > 4y\right\} &\subseteq \left\{\sup_{s\leq v\leq\tau} [\widehat{n}_{t,h,v,\tau}]\mathcal{K}(\widehat{\mu}_{t,h,v,\tau},\theta) > y\right\} \\ &\quad \cup \left\{\sup_{s\leq\tau} [\widehat{n}_{t,h,\tau}]\mathcal{K}(\widehat{\theta}_{t,h,\tau},\theta) > y\right\}.\end{aligned}$$

From Proposition I.14 and the previous inclusion, we obtain

$$\begin{aligned}&\mathbf{P}\left(\sup_{s\leq v\leq\tau} LR_{t,h}(v,\tau) > 4y\right) \\ \leq\ &\mathbf{P}\left(\sup_{s\leq v\leq\tau} [\widehat{n}_{t,h,v,\tau}]\mathcal{K}(\widehat{\mu}_{t,h,v,\tau},\theta) > y\right) + \mathbf{P}\left(\sup_{s\leq\tau} [\widehat{n}_{t,h,\tau}]\mathcal{K}(\widehat{\theta}_{t,h,\tau},\theta) > y\right) \\ \leq\ &2n^7 \exp\left(-\frac{y}{2} + \frac{d_{t,h,s,\theta}}{2}\right) + \frac{2}{n}.\end{aligned}$$

□

Lemma I.16. *If the sequences* $\{h_n\}$ *and* $\{\tau_n\}$ *are such that* $\bar{n}_{t,h_n,\tau_n} \to \infty$ *as* $n \to \infty$, *then* $\widehat{n}_{t,h_n,\tau_n} \overset{\mathbf{P}}{\asymp} \bar{n}_{t,h_n,\tau_n}$ *as* $n \to \infty$, *where* $\bar{n}_{t,h,\tau}$ *is defined by (II.7). Moreover,* $[\widehat{n}_{t,h_n,\tau_n}] \overset{\mathbf{P}}{\asymp} \bar{n}_{t,h_n,\tau_n}$ *as* $n \to \infty$.

Proof. By Chebyshev's exponential inequality, for any $u > 0$ and $\epsilon \in (0,1)$,

$$\mathbf{P}\left(\frac{\widehat{n}_{t,h_n,\tau_n}}{\bar{n}_{t,h_n,\tau_n}} < 1-\epsilon\right) \leq \exp\left(u(1-\epsilon)\bar{n}_{t,h_n,\tau_n} + \log \mathbf{E}\left(e^{-u\widehat{n}_{t,h_n,\tau_n}}\right)\right). \tag{I.41}$$

Applying Hölder's inequality, we have

$$\begin{aligned}\log \mathbf{E}e^{-u\widehat{n}_{t,h_n,\tau_n}} &= \sum_{i=1}^{n} \log \mathbf{E}\left(\left(e^{-u1_{\{X_{t_i}>\tau_n\}}}\right)^{W_{t,h_n}(t_i)}\right) \\ &\leq \sum_{i=1}^{n} W_{t,h_n}(t_i) \log \mathbf{E}\left(e^{-u1_{\{X_{t_i}>\tau_n\}}}\right).\end{aligned}$$

Using the fact that

$$\log \mathbf{E}\left(e^{-u1_{\{X_{t_i}>\tau_n\}}}\right) = \log\left(1-(1-F_{X_{t_i}}(\tau_n))(1-e^{-u})\right) \leq -\left(1-F_{X_{t_i}}(\tau_n)\right)\left(1-e^{-u}\right)),$$

we have

$$u\bar{n}_{t,h_n,\tau_n} + \log \mathbf{E}\left(e^{-u\widehat{n}_{t,h_n,\tau_n}}\right) \leq \bar{n}_{t,h_n,\tau_n}(e^{-u}+u-1) \leq n_{t,h_n,\tau_n}\frac{u^2}{2}. \tag{I.42}$$

From (I.41) and (I.42) it follows that

$$\mathbf{P}\left(\frac{\widehat{n}_{t,h_n,\tau_n}}{\bar{n}_{t,h_n,\tau_n}} < 1-\epsilon\right) \leq \exp\left(\bar{n}_{t,h_n,\tau_n}(-u\epsilon + \frac{u^2}{2})\right). \tag{I.43}$$

In the same way, we have

$$\mathbf{P}\left(\frac{\widehat{n}_{t,h_n,\tau_n}}{\bar{n}_{t,h_n,\tau_n}} > 1+\epsilon\right) \leq \exp\left(\bar{n}_{t,h_n,\tau_n}\left(e^u - (1+\epsilon)u - 1\right)\right). \tag{I.44}$$

Note that, there exist $\delta > 0$ such that, for all $0 < u < \delta$ it holds $\frac{e^u-u-1}{u} < \epsilon$. Taking $u = \min\left\{\epsilon, \frac{\delta}{2}\right\}$, we obtain

$$(e^u - (1+\epsilon)u - 1) < 0, \text{ and } -u\epsilon + \frac{u^2}{2} < 0. \tag{I.45}$$

Then from (I.43), (I.44) and (I.45), we have

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\widehat{n}_{t,h_n,\tau_n}}{\bar{n}_{t,h_n,\tau_n}} < 1-\epsilon\right) = \lim_{n\to\infty} \mathbf{P}\left(\frac{\widehat{n}_{t,h_n,\tau_n}}{\bar{n}_{t,h_n,\tau_n}} > 1+\epsilon\right) = 0,$$

which is equivalent to

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{\widehat{n}_{t,h_n,\tau_n}}{\bar{n}_{t,h_n,\tau_n}} - 1\right| > \epsilon\right) = 0,$$

or $\widehat{n}_{t,h_n,\tau_n} \overset{\mathbf{P}}{\asymp} \bar{n}_{t,h_n,\tau_n}$ as $n \to \infty$. On the other hand, since $\bar{n}_{t,h_n,\tau_n} \to \infty$ as $n \to \infty$, we have

$$[\widehat{n}_{t,h_n,\tau_n}] \leq \widehat{n}_{t,h_n,\tau_n} \leq 2\,[\widehat{n}_{t,h_n,\tau_n}],$$

which implies that $[\widehat{n}_{t,h_n,\tau_n}] \overset{\mathbf{P}}{\asymp} \bar{n}_{t,h_n,\tau_n}$, as $n \to \infty$.This completes the proof. □

Lemma I.17. *For any given integer positive k_0 and any sequences $\{h_n\}$, $\{\tau_n\}$ satisfying $\bar{n}_{t,h_n,\tau_n} \to \infty$ as $n \to \infty$, it holds $\lim_{n\to\infty} \mathbf{P}(Y_{k_0} > \tau_n) = 1$.*

Proof. Since $\bar{n}_{t,h_n,\tau_n} \to \infty$ as $n \to \infty$, by Lemma I.16, there exist constants $C_1, C_2 > 0$ such that, as $n \to \infty$,

$$\mathbf{P}(C_1\bar{n}_{t,h_n,\tau_n} \leq \widehat{n}_{t,h_n,\tau_n} \leq C_2\bar{n}_{t,h_n,\tau_n}) \to 1,$$

and, for $M = k_0/C_1$, there exist $n_0 > 0$ such that, for any $n > n_0$,

$$\bar{n}_{t,h_n,\tau_n} > M.$$

So that, for any $n > n_0$,

$$\{C_1 \bar{n}_{t,h_n,\tau_n} \leq \widehat{n}_{t,h_n,\tau_n}\} \subseteq \{\widehat{n}_{t,h_n,\tau_n} > k_0\}.$$

Since $0 \leq W_{t,h_n}(t_i) \leq 1, i = 1, ..., n$, we have

$$\{\widehat{n}_{t,h_n,\tau_n} > k_0\} \subseteq \{Y_{k_0} > \tau_n\}.$$

It follows that

$$\mathbf{P}(Y_{k_0} > \tau_n) \to 1,$$

which ends the proof. □

2.8.1 Proof of Theorem I.1

Using the first inequality of Proposition I.13 with $s = v = \tau_n, h = h_n, \theta = \theta_n$ and $y = 4 \log n + d_{t,h_n,\tau_n,\theta_n}$, we have, as $n \to \infty$,

$$\mathcal{K}\left(\widehat{\theta}_{t,h_n,\tau_n}, \theta_n\right) = O_{\mathbf{P}}\left(\frac{1}{[\widehat{n}_{t,h_n,\tau_n}]}\left(4 \log n + d_{t,h_n,\tau_n,\theta_n}\right)\right).$$

Since by lemma I.16, $[\widehat{n}_{t,h_n,\tau_n}] \overset{\mathbf{P}}{\asymp} \bar{n}_{t,h_n,\tau_n}$, we have, as $n \to \infty$,

$$\mathcal{K}\left(\widehat{\theta}_{t,h_n,\tau_n}, \theta_n\right) = O_{\mathbf{P}}\left(\frac{1}{\bar{n}_{t,h_n,\tau_n}}\left(4 \log n + d_{t,h_n,\tau_n,\theta_n}\right)\right).$$

2.8.2 Proof of Proposition I.3

We fix some notations. For any distribution function F supported on the interval $[x_0, \infty)$, $x_0 \geq 0$ and having a strictly positive density f_F we define

$$\alpha_F(x) = \frac{1}{x \lambda_F(x)}, \quad x \geq x_0,$$

where $\lambda_F(x) = \frac{f_F(x)}{1-F(x)}$ is the hazard rate function corresponding to F. For $t \in [0, T_{\max}]$ we consider the neighborhood $U_{t,h} = [t-h, t+h]$ with $h > 0$. Define the distance

$$\rho_*(x, y) = \max\left\{\left|\log \frac{x}{y}\right|, \left|\frac{1}{x} - \frac{1}{y}\right|\right\}, \quad x, \ y > 0.$$

Without loss of generality, we can assume that $A_{max} > 1$. For any $t \in [0, T_{\max}]$, we shall determine two sequences $\{\tau_n\}$ and $\{h_n\}$ such that

$$\sup_{s\in U_{t,h_n}} \sup_{x\geq\tau_n} \rho_*(\alpha_{F_s,\tau_n}, \theta_{t,\tau_n})^2 = O\left(\frac{n_{t,h_n}}{n_{t,h_n}(1-F_t(\tau_n))}\right) \to 0 \text{ as } n \to \infty \tag{I.46}$$

and

$$\sup_{s\in U_{t,h_n}} \int_{\tau_n}^{\infty} \left(1+\log\frac{x}{\tau_n}\right)^2 \left(\frac{x}{\tau_n}\right)^{\epsilon_0} \frac{f_s(x)dx}{1-F_s(\tau_n)} \leq \epsilon_1, \text{for any } n \geq N\ , \tag{I.47}$$

where ϵ_0, ϵ_1 and N are some constants. To do this, let $0 < \epsilon_0 < \frac{1}{\gamma_{max}}$ and $t \in [0, T_{\max}]$. Denote

$$I_t(\tau) = \int_{\tau}^{\infty} \left(1+\log\frac{x}{\tau}\right)^2 \left(\frac{x}{\tau}\right)^{\epsilon_0} \frac{f_t(x)dx}{1-F_t(\tau)}, \tau \geq x_0.$$

We first prove that

$$I_t(\tau) \leq \epsilon_1, \tag{I.48}$$

for all $t \in [0, T_{\max}]$ and $\tau \geq N$ with some constants $\epsilon_1 > 0$ and $N > 0$. Since $\rho > 0$, $0 < \gamma_{min} \leq \gamma_t \leq \gamma_{max} < \infty$, $|r_t(x)| \leq A_t x^{-\frac{\rho}{\gamma_t}}$ and $|R_t(x)| \leq A_t x^{-\frac{\rho}{\gamma_t}}, x \geq x_0$, we have

$$A_t x^{-\frac{\rho}{\gamma_t}} \leq A_{max} x^{-\frac{\rho}{\gamma_{max}}} \leq \frac{1}{2} \tag{I.49}$$

for any $x \geq \max\left(x_0, (2A_{max})^{\frac{\gamma_{max}}{\rho}}\right) = N$. Hence,

$$\frac{1+r_t(x)}{1+R_t(\tau)} \leq 1 + \left|\frac{1+r_t(x)}{1+R_t(\tau)} - 1\right| \leq 1 + 4A\tau^{-\frac{\rho}{\gamma_{max}}} \leq 3, \tag{I.50}$$

for any $x \geq \tau \geq N$. Since $0 < \epsilon_0 < \frac{1}{\gamma_{max}}$, there exists an integer $k > 0$ such that $\epsilon_0 + \frac{1}{k} < \frac{1}{\gamma_{max}}$. Using (I.50) and $\log\frac{x}{\tau} \leq k\left(\frac{x}{\tau}\right)^{1/2k}$, for all $x \geq \tau$,we deduce, for any $\tau \geq N$,

$$\begin{aligned} I_t(\tau) &= \int_{\tau}^{\infty} \left(1+\log\frac{x}{\tau}\right)^2 \left(\frac{x}{\tau}\right)^{\epsilon_0-\frac{1}{\gamma_t}-1} \frac{(1+r_t(x))}{1+R_t(\tau)}\frac{dx}{\tau} \\ &\leq 3\int_{\tau}^{\infty} \left(1+\log\frac{x}{\tau}\right)^2 \left(\frac{x}{\tau}\right)^{\epsilon_0-\frac{1}{\gamma_{max}}-1} \frac{dx}{\tau} \\ &\leq 3\int_{\tau}^{\infty} 4k^2 \left(\frac{x}{\tau}\right)^{\epsilon_0+\frac{1}{k}-\frac{1}{\gamma_{max}}-1} \frac{dx}{\tau}. \end{aligned}$$

By taking $\epsilon_1 = \frac{12k^2}{\frac{1}{\gamma_{max}} - \epsilon_0 - \frac{1}{k}}$, the inequality (I.48) is satisfied.

Next, we prove that

$$\sup_{x \geq \tau} \rho_*(\alpha_{F_{s,\tau}(x)}, \theta_{t,\tau}) \leq C_\alpha \left(12 A_t \gamma_{max} \tau^{-\frac{\rho}{\gamma_t}} + 4 A_t \gamma_{max} \tau^{-\frac{\rho}{\gamma_s}} + L_\gamma h^\beta\right), \tag{I.51}$$

for any $s \in U_{t,h}$ and $\tau \geq N$ with some constant $C_\alpha > 0$. Indeed, we have

$$\alpha_{F_t}(x) = \left(\frac{1 + R_t(x)}{1 + r_t(x)}\right) \gamma_t.$$

From (I.49), for all $x \geq N$

$$\left|\frac{1 + R_t(x)}{1 + r_t(x)} - 1\right| = \frac{|R_t(x) - r_t(x)|}{1 + r_t(x)} \leq \frac{|R_t(x)| + |r_t(x)|}{1 + r_t(x)} \leq 4 A_t x^{-\frac{\rho}{\gamma_t}} \leq 2.$$

This implies for any x, $\tau \geq N$

$$\sup_{x \geq \tau} |\alpha_{F_t}(x) - \gamma_t| \leq 4 A_t \gamma_t \tau^{-\frac{\rho}{\gamma_t}}, \tag{I.52}$$

and

$$0 < \alpha_{min} \leq \alpha_{F_t}(x) \leq \alpha_{max} < \infty,$$

where $\alpha_{min} = \gamma_{min}$ and $\alpha_{max} = 3\gamma_{max}$. Integrating by parts in (I.22), we obtain

$$\begin{aligned} \theta_{t,\tau} &= \log \frac{x}{\tau} \frac{1 - F_t(x)}{1 - F_t(\tau)} \Big|_\tau^\infty + \int_\tau^\infty \frac{1 - F_t(x)}{1 - F_t(\tau)} \frac{dx}{x} \\ &= \int_\tau^\infty \left(\frac{x}{\tau}\right)^{-\frac{1}{\gamma_t} - 1} \frac{1 + R_t(x)}{1 + R_t(\tau)} \frac{dx}{\tau}. \end{aligned} \tag{I.53}$$

From (I.49), for any $x \geq N$

$$\left|\frac{1 + R_t(x)}{1 + R_t(\tau)} - 1\right| \leq 4 A_t \tau^{-\frac{\rho}{\gamma_t}} \leq 2. \tag{I.54}$$

Combining (I.53) and (I.54) gives

$$|\theta_{t,\tau} - \gamma_t| \leq \int_\tau^\infty \left(\frac{x}{\tau}\right)^{-\frac{1}{\gamma_t} - 1} \left|\frac{1 + R_t(x)}{1 + R_t(\tau)} - 1\right| \frac{dx}{\tau} \leq 4 A_t \gamma_t \tau^{-\frac{\rho}{\gamma_t}} \leq 2\gamma_t. \tag{I.55}$$

Therefore $\gamma_{min} \leq \theta_{t,\tau} \leq 3\gamma_{max}$. Using $\alpha_{F_t}(x) = \alpha_{F_{t,\tau}}(x)$ for any $x \geq \tau \geq x_0$, from (I.52) and (I.55) we have, for any $x \geq \tau \geq N$,

$$\begin{aligned}
\sup_{x\geq\tau} \rho_*(\alpha_{F_{t,\tau}}(x), \theta_{t,\tau}) &\leq C_\alpha \sup_{x\geq\tau} |\alpha_{F_{t,\tau}}(x) - \theta_{t,\tau}| \\
&\leq C_\alpha \left(\sup_{x\geq\tau} |\alpha_{F_t}(x) - \gamma_t| + |\theta_{t,\tau} - \gamma_t| \right) \\
&\leq 8 C_\alpha A_t \gamma_t \tau^{-\frac{\rho}{\gamma_t}},
\end{aligned}$$

where $C_\alpha = \max(\alpha_{min}^{-1}, \alpha_{min}^{-2})$. Taking into account that $\gamma_t \leq \gamma_{\max}$, we conclude

$$\sup_{x\geq\tau} \rho_*(\alpha_{F_{t,\tau}}(x), \theta_{t,\tau}) \leq 8 C_\alpha A_t \gamma_{max} \tau^{-\frac{\rho}{\gamma_t}}. \tag{I.56}$$

Since, for any $t, s \in [0, T_{\max}]$, $x \geq N$,

$$\begin{aligned}
|\alpha_{F_t}(x) - \alpha_{F_s}(x)| &= \left| \left(\frac{1+R_t(x)}{1+r_t(x)} - 1 \right) \gamma_t + (\gamma_t - \gamma_s) + \left(1 - \frac{1+R_s(x)}{1+r_s(x)} \right) \gamma_s \right| \\
&\leq \gamma_{max} \left| \frac{1+R_t(x)}{1+r_t(x)} - 1 \right| + |\gamma_t - \gamma_s| + \gamma_{max} \left| \frac{1+R_s(x)}{1+r_s(x)} - 1 \right| \\
&\leq 4 A_t \gamma_{max} \left(x^{-\frac{\rho}{\gamma_t}} + x^{-\frac{\rho}{\gamma_s}} \right) + L_\gamma |t-s|^\beta
\end{aligned}$$

and

$$\rho_*(\alpha_{F_t}(x), \alpha_{F_s}(x)) \leq C_\alpha |\alpha_{F_t}(x) - \alpha_{F_s}(x)|,$$

we obtain

$$\rho_*(\alpha_{F_t}(x), \alpha_{F_s}(x)) \leq C_\alpha \left(4 A_t \gamma_{max} (x^{-\frac{\rho}{\gamma_t}} + x^{-\frac{\rho}{\gamma_s}}) + L_\gamma |t-s|^\beta \right). \tag{I.57}$$

From (I.56) and (I.57), we deduce, for any $x \geq \tau \geq N$, $s \in U_{t,h}$,

$$\begin{aligned}
\rho_*(\alpha_{F_{s,\tau}(x)}, \alpha_{F_{t,\tau}(x)}) &= \rho_*(\alpha_{F_t(x)}, \alpha_{F_s(x)}) \\
&\leq C_\alpha \left(4 A_t \gamma_{max} (\tau^{-\frac{\rho}{\gamma_t}} + \tau^{-\frac{\rho}{\gamma_s}}) + L_\gamma h^\beta \right).
\end{aligned}$$

Therefore,

$$\begin{aligned}\sup_{x\geq\tau}\rho_*(\alpha_{F_{s,\tau}(x)},\theta_{t,\tau}) &\leq \sup_{x\geq\tau}\rho_*(\alpha_{F_{s,\tau}(x)},\alpha_{F_{t,\tau}(x)})+\sup_{x\geq\tau}\rho_*(\alpha_{F_{t,\tau}(x)},\theta_{t,\tau})\\ &\leq C_\alpha\left(12A_t\gamma_{max}\tau^{-\frac{\rho}{\gamma_t}}+4A_t\gamma_{max}\tau^{-\frac{\rho}{\gamma_s}}+L_\gamma h^\beta\right),\end{aligned}$$

which proves that (I.51) is satisfied.

From inequality (I.51), by taking the sequences $\{\tau_n\}$ and $\{h_n\}$ such that $\tau_n\to\infty$ and $h_n\to 0$ as $n\to\infty$, we have

$$\lim_{n\to\infty}\sup_{s\in U_{t,h_n}}\sup_{x\geq\tau_n}\rho_*(\alpha_{F_{s,\tau_n}(x)},\theta_{t,\tau_n})=0.$$

Hence, for any $0<\epsilon_0<\frac{1}{\gamma_{max}}$, there exists $n_0>0$ such that

$$\sup_{s\in U_{t,h_n}}\sup_{x\geq\tau_n}\rho_*(\alpha_{F_{s,\tau_n}(x)},\theta_{t,\tau_n})\leq\epsilon_0 \tag{I.58}$$

for any $n\geq n_0$. Moreover, from (I.48), the inequality (I.47) is satisfied with τ_n and h_n defined above. Therefore, by Proposition 8.6 in [15], it follows that, for any $n\geq n_0$, $s\in U_{t,h_n}$,

$$\chi^2(F_{s,\tau_n},G_{\tau_n,\theta_{t,\tau_n}})\leq C(\epsilon_0,\epsilon_1)\sup_{x\geq\tau_n}\rho_*^2(\alpha_{F_{s,\tau_n}(x)},\theta_{t,\tau_n}),$$

where $C(\epsilon_0,\epsilon_1)=\epsilon_1\exp(\epsilon_0)$. This implies

$$\begin{aligned}&\sup_{s\in U_{t,h_n}}\chi^2(F_{s,\tau_n},G_{\tau_n,\theta_{t,\tau_n}})\\ \leq\ & C(\epsilon_0,\epsilon_1)\max(\alpha_{min}^{-2},\alpha_{min}^{-4})\left(12A\gamma_{max}\tau_n^{-\frac{\rho}{\gamma_t}}(1+\frac{1}{3}\tau_n^{\frac{\rho L_\alpha}{\gamma_{min}^2}}h_n^\beta)+L_\gamma h_n^\beta\right)^2.\end{aligned}$$

We now determine the location τ_n and bandwidth h_n satisfying (I.46). The balance conditions are

$$h_n^{2\beta}\asymp\tau_n^{-\frac{2\rho}{\gamma_t}},$$

and

$$\tau_n^{-\frac{2\rho}{\gamma_t}} \asymp \frac{\log n_{t,h_n}}{n_{t,h_n}\tau_n^{-\frac{1}{\gamma_t}}\left(1+A\tau_n^{-\frac{\rho}{\gamma_t}}\right)}.$$

The optimal choice is given by

$$\tau_n \asymp \left(\frac{\log n_{t,h_n}}{n_{t,h_n}}\right)^{\frac{-\gamma_t}{1+2\rho}}.$$

Taking into account that $n_{t,h_n} \asymp 2nh_n$, we obtain as $n \to \infty$

$$\begin{aligned} h_n &\asymp \left(\frac{\log n}{n}\right)^{\frac{1}{1+\beta(2+\rho^{-1})}}, \\ \tau_n &\asymp \left(\frac{\log n}{n}\right)^{\frac{-\gamma_t\beta\rho^{-1}}{1+\beta(2+\rho^{-1})}}, \\ \frac{\log n_{t,h_n}}{n_{t,h_n}(1-F_t(\tau_n))} &\asymp \left(\frac{\log n}{n}\right)^{\frac{2\beta}{1+\beta(2+\rho^{-1})}}, \end{aligned}$$

and

$$\chi^2(F_{s,\tau_n}, G_{\tau_n,\theta_{t,\tau_n}}) = O\left(\frac{\log n_{t,h_n}}{n_{t,h_n}(1-F_t(\tau_n))}\right)$$

uniformly in $s \in U_{t,h_n}$.This implies, as $n \to \infty$,

$$\sup_{s\in U_{t,h_n}} \chi^2(F_{s,\tau_n}, G_{\tau_n,\theta_{t,\tau_n}}) = O\left(\frac{\log n_{t,h_n}}{n_{t,h_n}(1-F_t(\tau_n))}\right).$$

Taking into account that, as $n \to \infty$,

$$\bar{n}_{t,h_n,\tau_n} = \sum_{t_i\in U_{t,h_n}} W_{t,h_n}(t_i)(1-F_{t_i}(\tau_n)) \asymp n_{t,h_n}(1-F_t(\tau_n)), \tag{I.59}$$

completes the proof.

2.8.3 Proof of Proposition I.5

We shall determine two sequences $\{\tau_n\}$ and $\{h_n\}$ satisfying $\tau_n \geq x_0$, $\tau_n \to \infty$, $h_n \to 0$ and

$$\sup_{x \geq \tau_n} \rho_*(\alpha_{F_s,\tau_n}(x), \theta_{t,\tau_n}) = o(1), \tag{I.60}$$

$$\int_{\tau_n}^{\infty} \left(1 + \log \frac{x}{\tau_n}\right)^2 \left(\frac{x}{\tau_n}\right)^{\epsilon_0(n)} F_{s,\tau_n}(dx) = O(1) \tag{I.61}$$

as $n \to \infty$, uniformly in $s \in U_{t,h_n}$.

We prove first (I.60). Indeed, by straightforward calculations, we have

$$\alpha_{F_t}(x) = \alpha(x, \gamma_t, \delta_t) = \frac{Cx^{-1/\gamma_t} + (1-C)x^{-1/\delta_t}}{\gamma_t^{-1}Cx^{-1/\gamma_t} + \delta_t^{-1}(1-C)x^{-1/\delta_t}}$$

and

$$\theta_{t,\tau} = \theta(\tau, \gamma_t, \delta_t) = \frac{\gamma_t C\tau^{-1/\gamma_t} + \delta_t(1-C)\tau^{-1/\delta_t}}{C\tau^{-1/\gamma_t} + (1-C)\tau^{-1/\delta_t}}.$$

It easy to see that there exist a constant $N > 0$ such that, for any $x \geq N$,

$$\left|\frac{\partial \alpha}{\partial \gamma_t}(x, \gamma_t, \delta_t)\right| \leq C_1(t) \text{ and } \left|\frac{\partial \alpha}{\partial \delta_t}(x, \gamma_t, \delta_t)\right| \leq C_1(t),$$

uniformly in γ_t and δ_t with some constant $C_1(t)$ depending on t. Therefore, by Taylor's expansion,

$$\begin{aligned}\alpha_{F_t}(x) - \alpha_{F_s}(x) &= \frac{\partial \alpha}{\partial \gamma_t}(x, \gamma_s + c(\gamma_t - \gamma_s), \delta_s + c(\delta_t - \delta_s))(\gamma_t - \gamma_s) + \\ &+ \frac{\partial \alpha}{\partial \delta_t}(x, \gamma_s + c(\gamma_t - \gamma_s), \delta_s + c(\delta_t - \delta_s))(\delta_t - \delta_s),\end{aligned}$$

where $c \in (0,1)$. So that, for any $x \geq N$,

$$|\alpha_{F_t}(x) - \alpha_{F_s}(x)| \leq L_\alpha |t-s|^\beta,$$

where $L_\alpha = 2C_1(t)L$.

Let $h > 0$. Since $\rho_*(\alpha_{F_t}(x), \alpha_{F_s}(x)) \leq \max(\delta_0^{-1}, \delta_0^{-2})|\alpha_{F_t}(x) - \alpha_{F_s}(x)|$, we have, for any

$s \in U_{t,h}$, $x \geq N$,

$$\rho_*(\alpha_{F_t}(x), \alpha_{F_s}(x)) \leq \max(\delta_0^{-1}, \delta_0^{-2}) L_\alpha h^\beta,$$

and, for any $\tau \geq N$,

$$\begin{aligned}
\rho_*(\alpha_{F_{s,\tau}}(x), \theta_{t,\tau}) &\leq \rho_*(\alpha_{F_{s,\tau}}(x), \alpha_{F_{t,\tau}}(x)) + \rho_*(\alpha_{F_{t,\tau}}(x), \theta_{t,\tau}) \\
&= \rho_*(\alpha_{F_s}(x), \alpha_{F_t}(x)) + \rho_*(\alpha_{F_{t,\tau}}(x), \theta_{t,\tau}) \\
&\leq \max(\delta_0^{-1}, \delta_0^{-2}) L_\alpha h^\beta + \rho_*(\alpha_{F_{t,\tau}}(x), \theta_{t,\tau}).
\end{aligned}$$

It follows that, for any $\tau \geq N$,

$$\sup_{x \geq \tau} \rho_*(\alpha_{F_{s,\tau}}(x), \theta_{t,\tau}) \leq \max(\delta_0^{-1}, \delta_0^{-2}) L_\alpha h^\beta + \sup_{x \geq \tau} \rho_*(\alpha_{F_{t,\tau}}(x), \theta_{t,\tau}),$$

uniformly in $s \in U_{t,h}$. On the other hand, in the same way as in Section 2.8.2 (cf. bound (8.14)), there exists a constant $C_2(t) > 0$ such that for any sequence (τ_n) satisfying $\tau_n \geq x_0$, we have

$$\sup_{x \geq \tau_n} \rho_*(\alpha_{F_{t,\tau_n}}(x), \theta_{t,\tau_n}) \leq C_2(t) \tau_n^{\frac{1}{\gamma_t} - \frac{1}{\delta_t}}.$$

By choosing $\tau_n \to \infty$ and $h = h_n$, where $\{h_n\}$ is a sequence satisfying $h_n \to 0$, we obtain

$$\sup_{x \geq \tau_n} \rho_*(\alpha_{F_{s,\tau_n}}(x), \theta_{t,\tau_n}) \leq \max(\delta_0^{-1}, \delta_0^{-2}) L_\alpha h_n^\beta + C_2(t) \tau_n^{\frac{1}{\gamma_t} - \frac{1}{\delta_t}} = \epsilon_0(n). \tag{I.62}$$

From this, we obtain (I.60).

We now prove (I.61). Denote $I_s(n) = \int_{\tau_n}^\infty \left(1 + \log \frac{x}{\tau_n}\right)^2 \left(\frac{x}{\tau_n}\right)^{\epsilon_0(n)} F_{s,\tau_n}(dx)$. As $\epsilon_0(n) \to 0$ as $n \to \infty$, there exist $n_0 > 0$ such that, for any $n > n_0$, $\epsilon_0(n) < \gamma_{\max}^{-1}$. By straightforward calculations

$$I_s(n) = \frac{C\gamma_s^{-1}}{C + (1-C)\tau_n^{\frac{1}{\gamma_t} - \frac{1}{\delta_t}}} g(\gamma_s^{-1} - \epsilon_0(n)) + \frac{(1-C)\delta_s^{-1}\tau_n^{\frac{1}{\gamma_t} - \frac{1}{\delta_t}}}{C + (1-C)\tau_n^{\frac{1}{\gamma_t} - \frac{1}{\delta_t}}} g(\delta_s^{-1} - \epsilon_0(n)),$$

where $g(x) = \frac{1}{x} + \frac{2}{x^2} + \frac{2}{x^3}, x \neq 0$ and $n \geq n_0$. It easy to see that, for any $n \geq n_0$ and $s \in U_{t,h_n}$,

$$I_s(n) \leq 2\delta_{\min}^{-1} g(\gamma_{\max}^{-1} - \epsilon_0(n)) = \epsilon_1(n).$$

This implies (I.61).

Combining (I.60) and (I.61), from Proposition 8.6 in [15], we have

$$\chi^2(F_{s,\tau_n}, G_{\tau_n,\theta_{t,\tau_n}}) \leq \epsilon_1(n) \exp(\epsilon_0(n)) \sup_{x \geq \tau_n} \rho_*(\alpha_{F_{s,\tau_n}}(x), \theta_{t,\tau_n}), \tag{I.63}$$

uniformly in $s \in U_{t,h_n}$, for all $n \geq n_0$. Note that $\epsilon_1(n) \exp(\epsilon_0(n)) \leq C_3$ for large n, say $n \geq n_1$, where n_1 is a constant. From (I.62) and (I.63), we obtain

$$\sup_{s \in U_{t,h_n}} \chi^2(F_{s,\tau_n}, G_{\tau_n,\theta_{t,\tau_n}}) \leq C_3 \left(\max(\delta_0^{-1}, \delta_0^{-2}) L_\alpha h_n^\beta + C_2(t) \tau_n^{\frac{1}{\gamma_t} - \frac{1}{\delta_t}} \right)^2$$

for any $n \geq \max\{n_0, n_1\}$. From this, we have the balance conditions for determining the oracle location τ_n, and oracle bandwidth h_n :

$$\tau_n^{\frac{2}{\gamma_t} - \frac{2}{\delta_t}} \asymp \frac{\log n_{t,h_n}}{n_{t,h_n}(C\tau_n^{-1/\gamma_t} + (1-C)\tau_n^{-1/\delta_t})} \qquad \text{and} \qquad h_n^{2\beta} \asymp \tau_n^{\frac{2}{\gamma_t} - \frac{2}{\delta_t}}. \tag{I.64}$$

The optimal choice is given by $\tau_n \asymp \left(\frac{n_{t,h_n}}{\log n_{t,h_n}}\right)^{\gamma_t \delta_t / (2\gamma_t - \delta_t)}$, where $\delta_t = \frac{\gamma_t}{1+\nu_t}$. Taking into account (I.64) with $n_{t,h_n} \asymp 2nh_n$, we obtain, as $n \to \infty$,

$$h_n \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{1+\beta(2+\nu_t^{-1})}}, \qquad \tau_n \asymp \left(\frac{\log n}{n}\right)^{\frac{-\beta\gamma_t\nu_t^{-1}}{1+\beta(2+\nu_t^{-1})}},$$

$$\frac{\log n_{t,h_n}}{n_{t,h_n}(1 - F_t(\tau_n))} = O\left(\left(\frac{\log n}{n}\right)^{\frac{2\beta}{1+\beta(2+\nu_t^{-1})}}\right),$$

and

$$\chi^2(F_{s,\tau_n}, G_{\tau_n,\theta_{t,\tau_n}}) = O\left(\frac{\log n_{t,h_n}}{n_{t,h_n}(1 - F_t(\tau_n))}\right), \text{ uniformly in } s \in U_{t,h_n}.$$

This implies, as $n \to \infty$,

$$\sup_{s \in U_{t,h_n}} \chi^2(F_{s,\tau_n}, G_{\tau_n,\theta_{t,\tau_n}}) = O\left(\frac{\log n_{t,h_n}}{n_{t,h_n}(1-F_t(\tau_n))}\right).$$

We complete the proof using (I.59).

2.8.4 Proof of Theorem I.7

First we prove that there exists c^* such that, for n sufficiently large,

$$\mathbf{P}\left(\sup_{\tau_n \le v \le \tau} \mathrm{LR}_{t,h_n}(v,\tau) > c^* \log n\right) \le \frac{4}{n}. \tag{I.65}$$

By Proposition I.15, we have, for any $y > 0$,

$$\mathbf{P}\left(\sup_{\tau_n \le v \le \tau} \mathrm{LR}_{t,h_n}(v,\tau) > 4y\right) \le 2n^7 \exp\left(-\frac{y}{2} + \frac{d_{t,h_n,\tau_n,\theta_n}}{2}\right) + \frac{2}{n},$$

where $d_{t,h_n,\tau_n,\theta_n} = \sum_{i=1}^n W_{t,h_n}(t_i)\chi^2(F_{t_i}, F_{t_i,\tau_n,\theta_n})$. Letting $y = 16\log n + \frac{1}{2}d_{t,h_n,\tau_n,\theta_n}$, we obtain

$$\mathbf{P}\left(\sup_{\tau_n \le v \le \tau} LR_{t,h_n}(v,\tau) > 4y\right) \le \frac{4}{n}.$$

Moreover, by condition C1, we can choose $4y < D = c^* \log n$, for some constant c^* and n sufficiently large, which implies (I.65). Now the assertion of the theorem comes from (I.65) and the inclusion

$$\left\{\sup_{X_{(k)} \ge \tau_n} \mathbf{Z}(X_{(k)}) > D\right\} \subseteq \left\{\sup_{\tau_n \le v \le \tau} \mathrm{LR}_{t,h_n}(v,\tau) > D\right\}.$$

2.8.5 Proof of Theorem I.8

We use the notations $\widehat{n}_r = \widehat{n}_{t,h_n,Y_r}$, $\widehat{\theta}_r = \widehat{\theta}_{t,h_n,Y_r}$ and

$$\mathbf{Z}(\tau_n) = \max_{\delta' \widehat{n}_{t,h_n,\tau_n} \le \widehat{n}_l \le (1-\delta'')\widehat{n}_{t,h_n,\tau_n}} LR_{t,h_n}(\tau_n, Y_l),$$

for $r = 2, ..., n$. Let

$$\Omega_{t,h_n,\tau_n} = \bigcap_{Y_r \geq \tau_n} \{\mathbf{Z}(Y_r) \leq D\} \bigcap \{\mathbf{Z}(\tau_n) \leq D\}$$

and

$$\Omega^*_{t,h_n,\tau_n} = \Omega_{t,h_n,\tau_n} \bigcap \{\mathbf{Z}(Y_{k_0}) \leq D\}.$$

Obviously, we have

$$\Omega_{t,h_n,\tau_n} \cap \{Y_{k_0} \geq \tau_n\} \subseteq \Omega^*_{t,h_n,\tau_n}.$$

Since $\Omega^c_{t,h_n,\tau_n} \subseteq \left\{\sup_{\tau_n \leq v \leq \tau} LR_{t,h_n}(v,\tau) > D\right\}$, by Theorem I.7, it follows that

$$\lim_{n\to\infty} \left(\Omega^c_{t,h_n,\tau_n}\right) = 0.$$

From this and Lemma I.17, we obtain,

$$\lim_{n\to\infty} \left(\Omega^*_{t,h_n,\tau_n}\right) = 1 \tag{I.66}$$

By the definition of $\widehat{k}$, on the set Ω^*_{t,h_n,τ_n}, it holds $\widehat{n}_{\widehat{k}-1} \geq \widehat{n}_{t,h_n,\tau_n}$.

First we compare $\widehat{\theta}_{\widehat{k}-1}$ and $\widehat{\theta}_{t,h_n,\tau_n}$. To this end define the sequence of natural numbers $m_i, i = 0, ..., i^*$, such that $m_0 = \widehat{k} - 1$ and $\delta'\widehat{n}_{m_{i-1}} \leq \widehat{n}_{m_i} \leq \frac{1}{2}\widehat{n}_{m_{i-1}} \leq (1-\delta'')\widehat{n}_{m_{i-1}}$, for $i = 1, ..., i^*$, where i^* such that $\delta'\widehat{n}_{m_{i^*}} \leq \widehat{n}_{t,h_n,\tau_n} \leq (1-\delta'')\widehat{n}_{m_{i^*}}$. Denote $\widehat{n}_{m_{i^*+1}} = \widehat{n}_{t,h_n,\tau_n}$ and $\widehat{\theta}_{m_{i^*+1}} = \widehat{\theta}_{t,h_n,\tau_n}$. Since, on the set Ω^*_{t,h_n,τ_n},

$$\mathbf{Z}(\tau_n) \leq D = c^* \log n$$

and

$$\mathbf{Z}(Y_r) \leq D = c^* \log n, \quad \text{for} \quad k_0 \leq r \leq \widehat{k} - 1,$$

by (I.29) with $s = Y_{m_{i-1}} \leq \tau = Y_{m_i}$, we have

$$\widehat{n}_{m_i} \mathcal{K}(\widehat{\theta}_{m_i}, \widehat{\theta}_{m_{i-1}}) \leq LR_{t,h_n}(s,\tau) \leq D, \ i = 1, ..., i^*.$$

In the same way, with $s = Y_{m_{i^*}} \leq \tau = \tau_n$, we have

$$\widehat{n}_{t,h_n,\tau_n} \mathcal{K}(\widehat{\theta}_{t,h_n,\tau_n}, \widehat{\theta}_{m_{i^*}}) \leq LR_{t,h_n}(s,\tau) \leq D.$$

This, in turn, implies

$$\sum_{i=1}^{i^*+1} \sqrt{\mathcal{K}(\widehat{\theta}_{m_i}, \widehat{\theta}_{m_{i-1}})} \leq D^{1/2} \sum_{i=1}^{i^*+1} \widehat{n}_{m_i}^{-1/2}.$$

Taking into account that $\widehat{n}_{m_i} \leq \frac{1}{2}\widehat{n}_{m_{i-1}}$, for $i = 1, ..., i^* + 1$, we obtain

$$\sum_{i=1}^{i^*+1} \widehat{n}_{m_i}^{-1/2} \leq \widehat{n}_{m_{i^*+1}}^{-1/2} \sum_{i=1}^{i^*+1} 2^{-(i^*-i+1)/2} \leq (2+\sqrt{2})\widehat{n}_{m_{i^*+1}}^{-1/2}.$$

According to Lemma 8.1 and 8.2 in [15], for n sufficiently large, it holds

$$\sqrt{\mathcal{K}(\widehat{\theta}_{m_{i^*+1}}, \widehat{\theta}_{m_0})} \leq \frac{3}{2} \sum_{i=1}^{i^*+1} \sqrt{\mathcal{K}(\widehat{\theta}_{m_i}, \widehat{\theta}_{m_{i-1}})} \leq \frac{3}{2}(2+\sqrt{2}) D^{1/2} \widehat{n}_{m_{i^*+1}}^{-1/2},$$

and

$$\begin{aligned} \sqrt{\mathcal{K}(\widehat{\theta}_{\widehat{k}-1}, \widehat{\theta}_{t,h_n,\tau_n})} &= \sqrt{\mathcal{K}(\widehat{\theta}_{m_0}, \widehat{\theta}_{m_{i^*+1}})} \\ &\leq \frac{9}{4}(2+\sqrt{2}) D^{1/2} \widehat{n}_{m_{i^*+1}}^{-1/2} \\ &= \frac{9}{4}(2+\sqrt{2}) \sqrt{c^* \frac{\log n}{\widehat{n}_{t,h_n,\tau_n}}}. \end{aligned} \tag{I.67}$$

Now we shall compare $\widehat{\theta}_{\widehat{k}-1} = \widehat{\theta}_{m_0}$ and $\widehat{\theta}_{\widehat{l}}$. Recall that $\widehat{l}$ satisfies

$$\delta' \widehat{n}_{t,h_n,\tau_n} \leq \delta' \widehat{n}_{\widehat{k}-1} \leq \widehat{n}_{\widehat{l}} \leq (1-\delta'') \widehat{n}_{\widehat{k}-1}.$$

Since, on the set Ω^*_{t,h_n,τ_n}, we have $LR_{t,h_n}(Y_{\widehat{k}-1}, Y_{\widehat{l}}) = \mathbf{Z}(Y_{\widehat{k}-1}) \leq D = c^* \log n$, it follows that

$$\sqrt{\mathcal{K}(\widehat{\theta}_{\widehat{l}}, \widehat{\theta}_{\widehat{k}-1})} \leq \sqrt{c^* \frac{\log n}{\widehat{n}_{\widehat{l}}}} \leq \sqrt{\frac{c^*}{\delta'} \frac{\log n}{\widehat{n}_{t,h_n,\tau_n}}}. \tag{I.68}$$

Combining (I.67) and (I.68), by Lemma 8.2 in [15], it follows that, on the set Ω^*_{t,h_n,τ_n},

$$\sqrt{\mathcal{K}(\widehat{\theta}_{\widehat{l}}, \widehat{\theta}_{t,h_n,\tau_n})} \leq \sqrt{cc^* \frac{\log n}{\widehat{n}_{t,h_n,\tau_n}}},$$

where c is a positive constant. Taking into account (I.66), we have

$$\mathbf{P}\left(\mathcal{K}(\widehat{\theta}_{\widehat{l}}, \widehat{\theta}_{t,h_n,\tau_n}) \leq cc^* \frac{\log n}{\widehat{n}_{t,h_n,\tau_n}}\right) \to 1 \quad \text{as} \quad n \to \infty.$$

Hence, by Lemma I.16, we obtain

$$\mathbf{P}\left(\mathcal{K}(\widehat{\theta}_{\widehat{l}}, \widehat{\theta}_{t,h_n,\tau_n}) \leq cc^* \frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right) \to 1 \quad \text{as} \quad n \to \infty.$$

The result follows.

2.8.6 Proof of Theorem I.10

From the decomposition

$$\mathcal{K}(F_{t,\tau}, G_{\tau,\theta}) = \mathcal{K}(F_{t,\tau}, G_{\tau,\theta_{t,\tau}}) + \int_\tau^\infty \log \frac{dG_{\tau,\theta_{t,\tau}}}{dG_{\tau,\theta}} dF_{t,\tau}$$

and the identity $\int_\tau^\infty \log \frac{dG_{\tau,\theta_{t,\tau}}}{dG_{\tau,\theta}} dF_{t,\tau} = \mathcal{K}(\theta_{t,\tau}, \theta)$, we have, for any $\theta > 0$ and any $\tau \geq x_0$,

$$\mathcal{K}(F_{t,\tau}, G_{\tau,\theta}) = \mathcal{K}(F_{t,\tau}, G_{\tau,\theta_{t,\tau}}) + \mathcal{K}(\theta_{t,\tau}, \theta). \tag{I.69}$$

Using (I.69) with $\tau = \tau_n, h = h_n$ and $\theta = \widehat{\theta}_{t,h_n,\widehat{\tau}_n}$, we have

$$\mathcal{K}(F_{t,\tau_n}, G_{\tau_n,\widehat{\theta}_{t,h_n,\widehat{\tau}_n}}) = \mathcal{K}(F_{t,\tau_n}, G_{\tau_n,\theta_{t,\tau_n}}) + \mathcal{K}(\theta_{t,\tau_n}, \widehat{\theta}_{t,h_n,\widehat{\tau}_n}). \tag{I.70}$$

From Theorem I.9 and Lemma 8.1 in [15], we have, for n sufficiently large,

$$\mathcal{K}(\theta_{t,\tau_n}, \widehat{\theta}_{t,h_n,\widehat{\tau}_n}) \leq \frac{9}{4}\mathcal{K}(\widehat{\theta}_{t,h_n,\widehat{\tau}_n}, \theta_{t,\tau_n}) = O_{\mathbf{P}}\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right) \tag{I.71}$$

as $n \to \infty$. Using condition (II.12) and the bounds

$$\mathcal{K}(F_{t,\tau_n}, G_{\tau_n,\theta_{t,\tau_n}}) \leq \log\left(1+\chi^2(F_{t,\tau_n}, G_{\tau_n,\theta_{t,\tau_n}})\right) \leq \chi^2(F_{t,\tau_n}, G_{\tau_n,\theta_{t,\tau_n}}),$$

we obtain

$$\mathcal{K}(F_{t,\tau_n}, G_{\tau_n,\theta_{t,\tau_n}}) = O\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right). \tag{I.72}$$

Combining (I.70), (I.71) and (I.72), it follows that as $n \to \infty$

$$\mathcal{K}\left(F_{t,\tau_n}, G_{\tau_n,\widehat{\theta}_{t,h_n,\widehat{\tau_n}}}\right) = O_{\mathbf{P}}\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right).$$

Chapitre I. Estimation non paramétrique des probabilités conditionnelles d'évènements rares et des quantiles extrêmes conditionnels

Chapitre II

Détermination simultannée du seuil τ et de la taille de la fenêtre h par une méthode adaptative

1 Introduction

Dans ce Chapitre, nous proposons une approche adaptative pour choisir la taille de la fenêtre h avec un seuil τ fixé. Nous démontrons la convergence de l'estimateur adaptatif obtenu et nous donnons également une méthode pour choisir simultanément la taille de la fenêtre h et le seuil τ.

Comme dans le Chapitre I, nous considérons $F_t(x) = P(X \leq x|T = t)$ la fonction de répartition conditionnelle d'une variable aléatoire X sachant $T = t \in [0, T_{\max}]$, où nous supposons que F_t appartient au domaine d'attraction de la loi de Fréchet. Nous observons les variables aléatoires indépendantes $X_{t_1}, \ldots, X_{t_n}$ associées aux temps $0 \leq t_1 < ... < t_n \leq T_{\max}$ où X_{t_i} a la distribution F_{t_i}.

L'approche utilisée ici pour le choix de la taille de la fenêtre h est basée sur une approximation paramétrique locale. En accord avec le paradigme de la vraisemblance locale, le paramètre θ_t peut être choisi comme étant constant sur un petit intervalle $[t-h, t+h]$, où le paramètre inconnu h doit être déterminé. Pour cette modélisation, la taille de la fenêtre est déterminée simultanément avec l'estimation du paramètre θ [44]. Le point crucial de la procédure qui consiste à choisir h est cependant différent de [44]. La détermination de h est effectuée ponctuellement pour chaque t. Nous commençons par tester l'hypothèse d'homogénéité de la fonction $(\theta_\cdot)$ sur

l'intervalle $[t-h, t+h]$ par un test de rapport de vraisemblance. Le paramètre h est déterminé par maximisation de la vraisemblance. La procédure s'écrit de la façon suivante :

Nous considérons un seuil $\tau \geq x_0$ déterministe. Soit une suite croissante de tailles de fenêtres $(\hbar_m)$. L'algorithme comporte deux étapes :

1. À partir de la plus petite taille de fenêtre, nous déterminons la plus grande taille $\hbar_{m^*}$ pour laquelle le modèle de Pareto n'est pas rejeté.

2. Parmi les tailles de fenêtres "acceptées" $\hbar_l$, nous choisissons la taille de fenêtre adaptative $\widehat{h}_t$ par maximum de vraisemblance. Pour exclure les valeurs proches de la taille de fenêtre rejetée, un terme de correction est rajouté.

Nous choisissons $\tau = \tau_n$, h_n et θ_n appelés respectivement seuil, fenêtre et paramètre Oracle et satisfaisant la condition II.10. Nous notons par $\widehat{h}_t$ le choix adaptatif de la taille de la fenêtre h. Le résultat principal de ce Chapitre est donné par le théorème suivant.

Théorème II.1. *Sous des conditions de régularités, nous avons, quand* $n \to \infty$,

$$\mathcal{K}(\widehat{\theta}_{t,\widehat{h}_t,\tau_n}, \theta_n) = O_{\mathbf{P}}\left(\frac{\log n}{\overline{n}_{t,h_n,\tau_n}}\right),$$

où $\overline{n}_{t,h,\tau} = \sum_{t_i \in I_{t,h}} (1 - F_{t_i}(\tau))$ *et* $I_{t,h} = [t-h, t+h] \cap [0, T_{\max}]$.

Nous donnons ensuite une procédure du choix simultané de la taille de la fenêtre h et du seuil τ. Pour cela, nous combinons les 2 procédures établies dans les paragraphes 2.4.1 et 2.2.2 : la première choisit τ quand la taille de la fenêtre h est fixée tandis que la seconde détermine la taille de la fenêtre h lorsque le seuil τ est fixé. La procédure pour un choix simultané de τ et h est semblable à celle du choix adaptatif de h, mais à chaque étape nous appliquons le choix adaptatif du seuil τ.

2 Article 2 : Non parametric adaptive estimation of conditional probabilities of rares events and extreme quantiles with simultaneous determination of the threshold and the bandwidth

Ce paragraphe concerne un projet d'article à soumettre.

2.1 Properties of the Oracle estimator

2.1.1 Formulation of the problem

We consider a pair of random variables (X, T), where X represents a quantity of interest and $T \in [0, T_{\max}]$ the time. Let $F_t(x) = P(X \leq x | T = t)$ be the conditional distribution of X given $T = t$ supported on the interval $[x_0, \infty)$, $x_0 \geq 0$ with a strictly positive density f_t. Let $x > x_0$ and $p \in (0, 1)$. As in Chapter I, the main aim is to provide a pointwise estimate of the tail probability $S_t(x) = 1 - F_t(x)$ and the extreme p-quantile $F_t^{-1}(p)$ processes on $[0, T_{\max}]$.

We assume that the distributions F_t are in the domain of attraction of the Fréchet distribution $\mathcal{D}(\Phi_{1/\theta_t}), \theta_t > 0$. By Fisher-Tippet-Gnedenko theorem (see Beirlant et al. [32]) this is equvalent to

$$\lim_{\tau \to \infty} \frac{1 - F_t(\lambda\tau)}{1 - F_t(\tau)} = \lambda^{-1/\theta_t}, \quad \text{for all } \lambda \geq 1. \tag{II.1}$$

For a fixed threshold $\tau \geq x_0$, let

$$F_{t,\tau}(x) = 1 - \frac{1 - F_t(x)}{1 - F_t(\tau)}, \quad x \in [\tau, \infty)$$

be the excess distribution function over the threshold τ. The condition (II.1) says that if x is large enough, $F_{t,\tau}(x)$ can be approximated by a Pareto distribution

$$G_{\tau,\theta_t}(x) = 1 - \left(\frac{x}{\tau}\right)^{-1/\theta_t}, \quad x \geq \tau \text{ as } \tau \to \infty. \tag{II.2}$$

The unknown parameter θ_t is the conditional tail index. This means that F_t can be approxi-

mated by the following semi-parametric model :

$$F_{t,\tau,\theta_t}(x) = \begin{cases} F_t(x) & \text{if } x \in [x_0, \tau], \\ 1 - (1 - F_t(\tau))(1 - G_{\tau,\theta_t}(x)) & \text{if } x > \tau \end{cases} \tag{II.3}$$

as the threshold τ converges to ∞. Thus, with the model which we shall use, we estimate F_t on the interval $[0, \tau]$ by the empirical distribution function, while on the interval $[\tau, \infty)$ we use the Pareto law with the parameter θ_t.

2.1.2 Estimation and some notations

We estimate the parameter θ_t using maximum likelihood method based on the observations in the window $I_{t,h} = [t-h, t+h] \cap [0, T_{\max}]$, where h is a bandwidth parameter. Our construction follows generally the same pattern as in Chapter 1 with the kernel $K(x) = 1_{[-1,1]}(x)$, where 1_A be the indicator of an event A, that is 1_A takes the values 1 when the condition A is verified and 0 otherwise. However, for these particular weights we found it useful to give the details here.

The maximum likelihood estimator of θ_t is given by

$$\widehat{\theta}_{t,h,\tau} = \arg\max_{\theta>0} \mathcal{L}_{t,h}(\tau, \theta),$$

where

$$\begin{aligned} \mathcal{L}_{t,h}(\tau,\theta) &= \sum_{t_i \in I_{t,h}} \log \frac{dF_{t,\tau,\theta}}{dx}(X_{t_i}) \\ &= \sum_{t_i \in I_{t,h}} 1_{\{X_{t_i} \le \tau\}} \log f_t(X_{t_i}) \\ &\quad + \sum_{t_i \in I_{t,h}} 1_{\{X_{t_i} > \tau\}} \log \left((1 - F_t(\tau)) \frac{1}{\tau\theta} \left(\frac{X_{t_i}}{\tau} \right)^{-\frac{1}{\theta}-1} \right). \end{aligned} \tag{II.4}$$

By elementary calculation, we obtain

$$\widehat{\theta}_{t,h,\tau} = \frac{1}{\widehat{n}_{t,h,\tau}} \sum_{t_i \in I_{t,h}} 1_{\{X_{t_i} > \tau\}} \log \left(\frac{X_{t_i}}{\tau} \right), \tag{II.5}$$

where

$$\widehat{n}_{t,h,\tau} = \sum_{t_i \in I_{t,h}} 1_{\{X_{t_i} > \tau\}} \tag{II.6}$$

is the number of the observations beyond the threshold τ.

2.1.3 Oracle estimator and its properties

It follows from the results in Chapter I that, under appropriate assumptions, the estimator $\widehat{\theta}_{t,h,\tau}$ converges to the Oracle parameter to be introduced below. To formulate the corresponding result, we introduce the necessary notations. Let

$$\mathcal{K}(P,Q) = \int \log \frac{dP}{dQ} dP$$

be the Kullback-Leibler entropy between two equivalent measures P and Q. For the Kullback-Leibler entropy between two Pareto distributions $G_{\tau,\theta'}$ and $G_{\tau,\theta}$ we use the notation :

$$\mathcal{K}(\theta',\theta) = \mathcal{K}(G_{\tau,\theta'}, G_{\tau,\theta}) = G\left(\frac{\theta'}{\theta} - 1\right)$$

where

$$G(x) = x - \log(x+1), \quad x > -1.$$

The χ^2 entropy between P and Q is defined by

$$\chi^2(P,Q) = \int \frac{dP}{dQ} dP - 1.$$

By Jensen's inequality we have $\chi^2(P,Q) \geq 0$. For any non-negative random variables A_n and B_n, the notation $A_n = O_{\mathbf{P}}(B_n)$ as $n \to \infty$ means that there exists a constant $c > 0$ such that $\mathbf{P}(A_n \leq cB_n) \to 1$ as $n \to \infty$. For any $t \in [0, T_{\max}]$ denote

$$\overline{n}_{t,h,\tau} = \sum_{t_i \in I_{t,h}} (1 - F_{t_i}(\tau)). \tag{II.7}$$

From the results of Chapter I (see Theorem I.1), we deduce the following theorem.

Theorem II.1. *Assume that* $\{\tau_n\}$ *and* $\{h_n\}$ *are two sequences such that* $\tau_n \geq x_0$ *and*

$$\overline{n}_{t,h_n,\tau_n} \to \infty \quad as \quad n \to \infty. \tag{II.8}$$

Then, for any sequence of positive numbers $\{\theta_n\}$, *we have as* $n \to \infty$,

$$\mathcal{K}\left(\widehat{\theta}_{t,h_n,\tau_n}, \theta_n\right) = O_{\mathbf{P}}\left(\frac{\log n}{\overline{n}_{t,h_n,\tau_n}} + \frac{1}{\overline{n}_{t,h_n,\tau_n}} \sum_{t_i \in I_{t,h}} \chi^2(F_{t_i}, F_{t_i,\tau_n,\theta_n})\right). \tag{II.9}$$

In the previous result $\{\theta_n\}$, $\{\tau_n\}$ and $\{h_n\}$ are some given sequences which should be determined in such a way that the right-hand side converges to 0 as fast as possible. They will generally depend on the unknown function F_t as the right-hand side of (II.9) depends on it. Therefore the importance of this result is merely theoretical, since in practical applications the distribution function F_t is not known. To obtain a computable estimator, we have to determine in an adaptive way the two sequences τ_n and h_n.

First, we introduce three Oracle sequences $\{\theta_n\}$, $\{\tau_n\}$ and $\{h_n\}$. The two terms in the right-hand side of (II.9) are the stochastic error and the modeling bias in the neighborhood of the time t. The optimal in order bound is obtained when the two terms, the stochastic error and the bias term in (II.9) are equal in order. This leads us the introduce the following small modeling bias condition.

C1. For any $t \in [0, T_{\max}]$ there exist sequences $\{\theta_n\}$, $\{\tau_n\}$ and $\{h_n\}$ (generally depending on t) such that

$$\sum_{t_i \in I_{t,h_n}} \chi^2(F_{t_i}, F_{t_i,\tau_n,\theta_n}) = O(\log n) \quad \text{as} \quad n \to \infty. \tag{II.10}$$

The parameters θ_n, τ_n and h_n satisfying condition C1 will be called respectively Oracle parameter, Oracle threshold and Oracle bandwidth. Condition C1 is a very general expression of the balance between the bias and the variance of the model. We can verify this condition when the family $(F_t)_{t\in[0,T_{\max}]}$ satisfies some regularity assumptions. For instance, C1 is satisfied for the Hall and mixture models in Sections 2.3.2 and 2.3.3.

Note that the bias term in the right-hand side of (II.9) can be bounded in the following way

$$\sum_{t_i \in I_{t,h}} \chi^2(F_{t_i}, F_{t_i,\tau_n,\theta_n}) \leq \sup_{t_i \in I_{t,h}} \chi^2(F_{t_i}, F_{t_i,\tau_n,\theta_n}), \tag{II.11}$$

where the right-hand side gives a more intuitive interpretation of the modeling bias introduced by our model.

The following condition, which obviously implies condition C1, will be used in the sequel.

C2. For any $t \in [0, T_{\max}]$ there exist sequences $\{\theta_n\}$, $\{\tau_n\}$ and $\{h_n\}$ (generally depending on t) such that

$$\sup_{s \in I_{t,h_n}} \chi^2(F_s, F_{s,\tau_n,\theta_n}) = O\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right). \tag{II.12}$$

The best approximation in (II.9) and (II.15) is attained when the sequences $\{\theta_n\}$, $\{\tau_n\}$ and $\{h_n\}$ are chosen such that as $n \to \infty$

$$\sum_{t_i \in I_{t,h_n}} \chi^2(F_{t_i}, F_{t_i,\tau_n,\theta_n}) \asymp \log n, \tag{II.13}$$

or

$$\sup_{s \in I_{t,h_n}} \chi^2(F_s, F_{s,\tau_n,\theta_n}) \asymp \log n. \tag{II.14}$$

From Theorem II.1, we have :

Theorem II.2. *Assume that the family $(F_t)_{t\in[0,T_{\max}]}$ satisfies condition C1 and*

$$\overline{n}_{t,h_n,\tau_n} \to \infty \quad as \quad n \to \infty.$$

Then, we have,

$$\mathcal{K}\left(\widehat{\theta}_{t,h_n,\tau_n}, \theta_n\right) = O_{\mathbf{P}}\left(\frac{\log n}{\overline{n}_{t,h_n,\tau_n}}\right) \quad as \quad n \to \infty. \tag{II.15}$$

Define is the best fitted Pareto parameter $\theta_{t,\tau}$, by minimizing the Kullback-Leibler entropy between the true excess distribution function $F_{t;\tau}$ and the fitted model Pareto model :

$$\theta_{t,\tau} = \arg\min_{\theta>0} \mathcal{K}(F_{t,\tau}, G_{\tau,\theta}) = \int_\tau^\infty \log \frac{x}{\tau} \frac{f_t(x)dx}{1 - F_t(\tau)}, \tau \geq x_0. \tag{II.16}$$

Theorem II.3. *Assume that the family $(F_t)_{t\in[0,T_{\max}]}$ satisfies condition C2 with $\theta_n = \theta_{t,\tau_n}$. Then, there exists a constant $c^* > 0$ in (II.30), such that as $n \to \infty$*

$$\mathcal{K}\left(F_{t,\tau_n}, G_{\tau_n,\widehat{\theta}_{t,h_n,\tau_n}}\right) = O_{\mathbf{P}}\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right). \tag{II.17}$$

Now we address the problem of adaptive selection of the parameters τ and h. The adaptive selection of the threshold τ for the non-adaptive estimator $\widehat{\theta}_{t,h,\tau}$, provided that $h = h_n$ is the Oracle bandwidth, has been considered in Chapter I (see Durrieu et al. [56]). To choose the bandwidth parameter h, in the same chapter, we have proposed a global approach based on cross validation.

In this chapter, we propose a local pointwise adaptive procedure to select the bandwidth parameter h. Both approaches appear to work well in applications, each having some advantages which will be discussed latter in Chapter III.

We end this section by giving the estimators of the distribution function F_t and its quantile function.

We recall that for any $t \in [0, T_{\max}]$, the empirical distribution function pertaining to the observations in the interval $I_{t,h}$ is

$$\widehat{F}_{t,h}(x) = \frac{1}{\sum_{j=1}^{n} 1_{I_{t,h}}(t_j)} \sum_{i=1}^{n} 1_{I_{t,h}}(t_i) 1_{\{X_{t_i}\le x\}}. \tag{II.18}$$

From (II.5) and (II.18), the distribution function $F_t(x)$ at time t is then estimated by

$$\widehat{F}_{t,h,\tau}(x) = \begin{cases} \widehat{F}_{t,h}(x) & \text{if } x \in [x_0, \tau], \\ 1 - \left(1 - \widehat{F}_{t,h}(\tau)\right)\left(\frac{x}{\tau}\right)^{-\frac{1}{\widehat{\theta}_{t,h,\tau}}} & \text{if } x > \tau, \end{cases} \tag{II.19}$$

which combines the empirical distribution function $\widehat{F}_{t,h}(x)$ and the fitted Pareto law.

For any $p \in (0,1)$, the estimator of the p-quantile of X_t is defined by

$$\widehat{q}_p(t) = \widehat{q}_p(t,h) \equiv \begin{cases} \widehat{F}^{-1}_{t,h}(p) & \text{if } \quad p < \widehat{p}_\tau, \\ \tau\left(\frac{1-\widehat{p}_\tau}{1-p}\right)^{\widehat{\theta}_{t,h,\tau}} & \text{otherwise,} \end{cases} \tag{II.20}$$

where $\widehat{p}_\tau = \widehat{F}_{t,h}(\tau)$.

2.2 Bandwidth selection and main results

We consider an approach for the bandwidth selection based on the local parametric approximation. According to the local likelihood paradigm, the parameter θ_t in (II.3) can be considered as constant in some presumably small window $[t-h, t+h]$, where the bandwidth parameter h has to be chosen. Under this type of modeling the bandwidth h will be determined simultaneously with the estimation of the parameter θ. We refer to Spokoiny [44] for a similar approach. Only the first part of our procedure based on the propagation principle is similar to that in Spokoiny [44]. The crucial part of the procedure which consists in the choice of the h is completely different.

The determination of the bandwidth h is performed pointwise in t. It starts with testing the assumption of the homogeneity of the parameter function $(\theta_\cdot)$ on the interval $[t-h, t+h]$ and is based on the likelihood maximization. It can be briefly described as follows.

Consider a deterministic threshold $\tau \geq x_0$ and an increasing sequence of bandwidths $(\hbar_m)$. The proposed procedure has two steps :

1. Starting from the smallest bandwidth, determine the largest $\hbar_{m^*}$ for which the Pareto model is not rejected.

2. Select among the accepted bandwidths $\hbar_l$, $l \leq m^*$ the adaptive bandwidth $\widehat{h}_t$ which maximizes the log likelihood, where, in order to exclude the values of $\hbar_l$ close to the rejected bandwidth, a correction term to the likelihood function is added.

Firstly, to state our convergence results, we choose $\tau = \tau_n$ to be the Oracle threshold. These results can be seen as a complement of those given in Chapter I where we show the convergence of the estimator $\widehat{\theta}_{t,h,\tau}$ with an adaptive choice of τ and $h = h_n$ being the Oracle bandwidth.

Latter on in Section 2.3, we propose a simultaneous data driven choice of the bandwidth h and the threshold τ.

The rest of this section is organized as follows. In Section 2.2.1, we describe a procedure for testing the homogeneity of the function θ_t on an interval which is adapted to our purposes. The results of this section are then applied in Section 2.2.2 to construct an adaptive procedure for choosing $\hbar_{m^*}$ and the adaptive bandwidth $\widehat{h}_t$.

2.2.1 Testing the homogeneity of the tail

Let the threshold parameter $\tau > x_0$ be fixed. Let the time $t \in [0, T_{\max}]$ be fixed and M be a positive integer. Consider a increasing sequence of bandwidths $0 < \hbar_1 < \hbar_2 < ... < \hbar_M$. An example used in our simulations is the uniform grid $\hbar_m = \hbar_1 + (m-1)\delta$, where $\hbar_1$ and δ are positive numbers, to be calibrated latter on. This sequence of bandwidths defines a family $\mathcal{I}_t = \{I_{t,\hbar_m} : m = 1, ..., M\}$ of nested intervals of the form $I_{t,\hbar_m} = [t - \hbar_m, t + \hbar_m] \cap [0, T_{\max}]$.

In this section we give a test to verify whether the family $(F_t)_{t\in[0,T_{\max}]}$ has a homogenous tail on the interval $I = I_{t,\hbar_m}$. More precisely, we will test whether the excess distribution function $F_{t_i,\tau}$ verifies the null hypothesis that it can be well approximated by a single Pareto model for all $t_i \in I$.

To perform the test, consider the family $\mathcal{J}_I$ of subintervals of $I = I_{t,\hbar_m}$ defined as follows :

$$\mathcal{J}_I = \mathcal{J}_{I_{t,\hbar_m}} = \{I_{t,\hbar_l} : \eta' \widehat{n}_{t,\hbar_m,\tau} \leq \widehat{n}_{t,\hbar_l,\tau} \leq (1-\eta'')\, \widehat{n}_{t,\hbar_m,\tau}\},$$

where $0 < \eta', \eta'' < 1/2$ are two real numbers. We introduce the following two hypotheses : the null hypothesis H_0 which consists in saying that

$$F_{t_i}(x) = F_{t_i,\tau,\theta}(x), \qquad \text{for all } \ t_i \in I = I_{t,\hbar_m} \tag{II.21}$$

and the alternative H_1 which consists in saying that there exist a subinterval $J = I_{t,\hbar_l} \in \mathcal{J}_I$

such that

$$F_{t_i}(x) = \begin{cases} F_{t_i,\tau,\theta'}(x) & \text{if } t_i \in J, \\ F_{t_i,\tau,\theta''}(x) & \text{if } t_i \in I \setminus J. \end{cases} \tag{II.22}$$

To test H_0 against H_1, we use the likelihood ratio test statistic

$$\begin{aligned} LR_{t,\tau}(\hbar_m, \hbar_l) &= \sup_{\theta'>0} \sum_{t_i \in I_{t,\hbar_l}} \log \frac{dF_{t_i,\tau,\theta'}}{dx}(X_{t_i}) + \sup_{\theta''>0} \sum_{t_i \in I_{t,\hbar_m} \setminus I_{t,\hbar_l}} \log \frac{dF_{t_i,\tau,\theta''}}{dx}(X_{t_i}) \\ &- \sup_{\theta>0} \sum_{t_i \in I_{t,\hbar_m}} \log \frac{dF_{t_i,\tau,\theta}}{dx}(X_{t_i}). \end{aligned}$$

By elementary calculation, we found that $LR_{t,\tau}(\hbar_m, \hbar_l)$ can be written in the form

$$LR_{t,\tau}(\hbar_m, \hbar_l) = \widehat{n}_{t,\hbar_l,\tau} \mathcal{K}(\widehat{\theta}_{t,\hbar_l,\tau}, \widehat{\theta}_{t,\hbar_m,\tau}) + \widehat{n}_{t,\hbar_m,\hbar_l,\tau} \mathcal{K}(\widehat{\theta}_{t,\hbar_m,\hbar_l,\tau}, \widehat{\theta}_{t,\hbar_m,\tau}), \tag{II.23}$$

where

$$\begin{aligned} \widehat{n}_{t,\hbar_m,\hbar_l,\tau} &= \widehat{n}_{t,\hbar_m,\tau} - \widehat{n}_{t,\hbar_l,\tau}, \\ \widehat{\theta}_{t,\hbar_m,\hbar_l,\tau} &= (\widehat{n}_{t,\hbar_m,\tau} \widehat{\theta}_{t,\hbar_m,\tau} - \widehat{n}_{t,\hbar_l,\tau} \widehat{\theta}_{t,\hbar_l,\tau}) / (\widehat{n}_{t,\hbar_m,\tau} - \widehat{n}_{t,\hbar_l,\tau}). \end{aligned}$$

We define the test statistic

$$\mathbb{Z}_{\hbar_m}(\tau) = \max_{I_{t,\hbar_l} \in \mathcal{J}_{I_{t,\hbar_m}}} LR_{t,\tau}(\hbar_m, \hbar_l). \tag{II.24}$$

The null hypothesis H_0 is rejected if

$$\mathbb{Z}_{\hbar_m}(\tau) > z,$$

where z is a critical value to be determined by simulations.

2.2.2 Adaptive selection of the bandwidth

The goal of this section is to determine a data driven interval of homogeneity $\widehat{I}$ in a given family of nested intervals. We shall test consecutively the homogeneity of the tails of the dis-

tributions $(F_t)_{t\in[0,T_{\max}]}$ on the intervals $I_{t,\hbar_m}$ in the set $\mathcal{I}_t$. The procedure starts by testing the homogeneity on the interval $I = I_{t,\hbar_{k_0}}$ against a non homogeneous alternative using the test formulated in the previous section. If the hypothesis of homogeneity is not rejected then we take the next larger interval $I = I_{t,\hbar_m}, k > k_0$ and test again for homogeneity on this larger interval. This procedure is continued until we detect the largest interval $I^* = I_{t,\hbar_{m^*}}$ on which the hypothesis H_0 is not rejected.

There are several possibilities for the choice of the "best" adaptive interval. One way, advised by the well known Lepski procedure [57], is to choose the intervall I^* as the adaptive interval. However, with this choice the test statistic has a value close to the critical value z which introduces a systematic bias when we fit the data on I^*.

A better fit can be obtained if we choose among the all accepted intervals using some criteria. For instance, in some settings, it would be useful to choose $I \in \mathcal{I}_t$ by maximizing the likelihood function similarly to the change point setting. However, by simulations we found, that the most appropriate choice is to maximise the likelihood in all subintervals I of I^* by taking into account a correction term which penalizes for getting close to I^*. To be more precise the adaptive interval I is defined as

$$\widehat{I} = \arg \max_{I_{t,\hbar_l}\in\mathcal{J}_{I_{t,\hbar_{m^*}}}} LR^{Pen}_{t,\tau}(\hbar_{m^*}, \hbar_l),$$

where

$$LR^{Pen}_{t,\tau}(\hbar_m, \hbar_l) = \sup_{\theta>0} \mathcal{L}_{t,\hbar_l}(\tau,\theta) - \mathcal{L}_{t,\hbar_l}(\tau, \widehat{\theta}_{t,h_k,\tau}). \tag{II.25}$$

The correction term $\mathcal{L}_{t,\hbar_l}(\tau, \widehat{\theta}_{t,h_k,\tau})$ in (II.25) is nothing else but the log-likelihood function computed on the interval $I_{t,\hbar_l}$ but with the parameter $\widehat{\theta}_{t,h_k,\tau}$ computed from the likelihood $\mathcal{L}_{t,h_k}(\tau,\theta)$ on the interval $I_{t,\hbar_m}$. The following formula shows that the penalized likelihood

$LR^{Pen}_{t,\tau}(\hbar_m, \hbar_l)$ has a value close to 0 when $\hbar_l$ becomes close to $\hbar_m$:

$$LR^{Pen}_{t,\tau}(\hbar_m, \hbar_l) = \widehat{n}_{t,\hbar_l,\tau}\mathcal{K}(\widehat{\theta}_{t,\hbar_l,\tau}, \widehat{\theta}_{t,\hbar_m,\tau}). \tag{II.26}$$

Note that $LR^{Pen}_{t,\tau}(\hbar_m, \hbar_l)$ is exactly the first term in the right-hand side of (II.23) with $I = I_{t,\hbar_m}$ and $J = I_{t,\hbar_l}, 1 \leq l < k \leq M$. The compact form (II.26) is obtained by some straightforward calculations from (II.25) and turns out to be useful to perform computations.

The procedure of the adaptive choice of h is as follows :

Step 1. Set $m = m_0$.

Step 2. Compute the test statistic (II.24) for testing the homogeneity :

$$\mathbb{Z}_{\hbar_m}(\tau) = \max_{\eta'\widehat{n}_{t,\hbar_m,\tau} \leq \widehat{n}_{t,\hbar_l,\tau} \leq (1-\eta'')\widehat{n}_{t,\hbar_m,\tau}} LR_{t,\tau}(\hbar_m, \hbar_l). \tag{II.27}$$

Step 3. If for $m \leq M$ a deviance from homogeneity is detected, i.e. $\mathbb{Z}_{\hbar_m}(\tau) > z$, set $m^* = m - 1$ and define the adaptive index by maximizing the penalized likelihood

$$\widehat{m} = \underset{\eta'\widehat{n}_{t,\hbar_{m^*},\tau} \leq \widehat{n}_{t,\hbar_l,\tau} \leq (1-\eta'')\widehat{n}_{t,\hbar_{m^*},\tau}}{\arg\max} LR^{\mathrm{Pen}}_{t,\tau}(\hbar_{m^*}, \hbar_l), \tag{II.28}$$

where $LR^{\mathrm{Pen}}_{t,\tau}$ is defined in (II.25). If for $m \leq M$ the test statistic computed above does not detect a deviation from the homogeneity, i.e. $\mathbb{Z}_{\hbar_m}(\tau) \leq z$, then : if $m < M$ increase m by 1 and return to Step 2 ; if $m = M$ set the adaptive index by $\widehat{m} = M$. Define the adaptive bandwidth by $\widehat{h}_t = \hbar_{\widehat{m}}$, the adaptive estimator by $\widehat{\theta}_{t,\widehat{h}_t,\tau}$ and exit the procedure.

We shall prove below the propagation property of the test statistic $\mathbb{Z}_{\hbar_m}(\tau)$, that is we show that under regularity conditions the adaptive bandwidth $\widehat{h}_t$ given by the adaptive selection procedure stated above is larger than the Oracle bandwidth h_n with probability close to 1, which means that the Oracle bandwidth h_n is selected by our procedure with hight probability.

Equivalently, we shall prove that if $0 < h \leq h_n$ then the test statistic

$$\mathbb{Z}_h(\tau_n) = \max_{\eta'\widehat{n}_{t,h,\tau_n} \leq \widehat{n}_{t,h_l,\tau_n} \leq (1-\eta'')\widehat{n}_{t,h,\tau_n}} LR_{t,\tau}(h, \hbar_l) \tag{II.29}$$

does not exceed the critical value

$$z = z(n, h_n) = c^* \log n_{t,h_n} \tag{II.30}$$

with high probability, for some constant $c^* > 0$. Note that (II.29) for $h = \hbar_m$ coincides with (II.29) for $\tau = \tau_n$.

Proposition II.4. *Assume that the family $(F_t)_{t\in[0,T_{\max}]}$ satisfies condition C1 with $h_n > \hbar_{k_0}$. Then, for any $t \in [0, T_{\max}]$, there exist a constant $c^* > 0$ in (II.30) such that for any $h > 0, h \in (\hbar_{k_0}, h_n]$*

$$\mathbb{P}\left(\mathbb{Z}_h(\tau_n) > c^* \log n_{t,h_n}\right) \leq \frac{2}{n_{t,h_n}^{3/2}}, \tag{II.31}$$

where h_n is the Oracle bandwidth from the condition C1. In particular, (II.31) holds for any $h = \hbar_k \in (\hbar_{k_0}, h_n]$.

The proof of this proposition is deferred to Section 2.4.

2.2.3 Convergence of the adaptive estimator

Now we state the main results of this chapter, which give a rate of convergence of the adaptive (in the bandwidth) estimator.

Our first result states that the adaptive (in the bandwidth) estimator and the non-adaptive estimators are close one to another.

Theorem II.5. *Assume that the family $(F_t)_{t\in[0,T_{\max}]}$ satisfies condition C1 with $h_n > \hbar_{k_0}$ and $\bar{n}_{t,h_n,\tau_n} \to \infty$ as $n \to \infty$. Then, for any $t \in [0, T_{\max}]$, there exist a constant $c^* > 0$ in (II.30) such that*

$$\mathcal{K}(\widehat{\theta}_{t,\widehat{h}_t,\tau_n}, \widehat{\theta}_{t,h_n,\tau_n}) = O_{\mathbb{P}}\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right) \text{ as } n \to \infty,$$

where $\bar{n}_{t,h_n,\tau_n} = \sum_{t_i \in I_{t,h_n}} (1 - F_{t_i}(\tau_n))$.

Proof. See Section 2.4.2. □

II.2 Article 2 : Non parametric adaptive estimation of conditional probabilities of rares events and extreme quantiles with simultaneous determination of the threshold and the bandwidth

The previous theorem allows to extend the results of the non adaptive setting to the adaptive one. Note that the rate of this approximation is the same as in Theorem II.2. Thus, our adaptive estimator $\widehat{\theta}_{t,\widehat{h}_t,\tau_n}$ approximates the unknown sequence θ_n with the same rate as the Oracle estimator $\widehat{\theta}_{t,h_n,\tau_n}$ does and therefore gives the same quality of estimation. The following theorem gives the corresponding result.

Theorem II.6. *Assume that the family $(F_t)_{t\in[0,T_{\max}]}$ satisfies condition C1. Then, there exists a constant $c^* > 0$ in (II.30), such that as $n \to \infty$,*

$$\mathcal{K}(\widehat{\theta}_{t,\widehat{h}_t,\tau_n}, \theta_n) = O_{\mathbf{P}}\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right).$$

Proof. The result is obtained by combining Theorems II.2 and II.5. □

In Theorems II.5 and II.6, the sequences h_n and τ_n are required to satisfy the condition C1.

Recall that the excess distribution function F_{t,τ_n} is estimated by the Pareto distribution $G_{\tau_n,\widehat{\theta}_{t,\widehat{h}_t,\tau_n}}$. We give now the rate of convergence of $G_{\tau_n,\widehat{\theta}_{t,\widehat{h}_t,\tau_n}}$ to F_{t,τ_n} in terms of the Kullback-Leibler divergence, where we replace condition C1 by the slightly stronger condition C2.

Theorem II.7. *Assume that the family $(F_t)_{t\in[0,T_{\max}]}$ satisfies condition C2 with $\theta_n = \theta_{t,\tau_n}$. Then, there exists a constant $c^* > 0$ in (II.30), such that as $n \to \infty$*

$$\mathcal{K}\left(F_{t,\tau_n}, G_{\tau_n,\widehat{\theta}_{t,\widehat{h}_t,\tau_n}}\right) = O_{\mathbf{P}}\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right).$$

Proof. The proof of this assertion is similar to that of Theorem I.10, Section 2.8.6. □

We shall discuss the result of Theorem II.6 and Theorem II.7, on two examples considered in Chapter I : the Hall model and the mixture of two Pareto distributions.

For these two examples, we have shown in Section 2.8.2 that condition C2 (and thus also condition C1) is satisfied. In the case of the Hall model (see Section 2.3.2), we obtain the following explicit rates of convergence.

Theorem II.8. *Assume that the family $(F_t)_{t\in[0,T_{\max}]}$ satisfies the assumptions of Proposition I.3. Then, there exists a constant $c^* > 0$ in (II.30), such that*

$$\sqrt{\mathcal{K}\left(\widehat{\theta}_{t,\widehat{h}_t,\tau_n}, \theta_{t,\tau_n}\right)} = O_{\mathbf{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{1+\beta(2+1/\rho)}}\right) \quad as \quad n \to \infty,$$

and

$$\sqrt{\mathcal{K}\left(F_{t,\tau_n}, G_{\tau_n,\widehat{\theta}_{t,\widehat{h}_t,\tau_n}}\right)} = O_{\mathbf{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{1+\beta(2+1/\rho)}}\right)$$

where the Oracle threshold τ_n satisfies

$$\tau_n \asymp \left(\frac{n}{\log n}\right)^{\frac{\gamma_t\beta/\rho}{1+\beta(2+1/\rho)}}.$$

Proof. This Theorem is a consequence of Proposition I.3, Theorems II.6 and II.7. □

For the mixture of two Pareto distributions (see Section 2.3.3), we obtain

Theorem II.9. *Assume that the family $(F_t)_{t\in[0,T_{\max}]}$ satisfies the assumptions of Proposition I.5. Then, there exists a constant $c^* > 0$ in (II.30), such that*

$$\sqrt{\mathcal{K}\left(\widehat{\theta}_{t,\widehat{h}_t,\tau_n}, \theta_{t,\tau_n}\right)} = O_{\mathbf{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{1+\beta(2+1/\rho_t)}}\right) \quad as \quad n \to \infty,$$

and

$$\sqrt{\mathcal{K}\left(F_{t,\tau_n}, G_{\tau_n,\widehat{\theta}_{t,\widehat{h}_t,\tau_n}}\right)} = O_{\mathbf{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{1+\beta(2+1/\rho_t)}}\right)$$

where the Oracle threshold τ_n satisfies

$$\tau_n \asymp \left(\frac{n}{\log n}\right)^{\frac{\gamma_t\beta/\rho_t}{1+\beta(2+1/\rho_t)}}.$$

Proof. This Theorem is a consequence of Proposition I.5, Theorems II.6 and II.7. □

2.3 Simultaneous choice of the threshold and bandwidth

The choice of the threshold τ and the bandwidth h have been discussed in Chapter I where we proposed an adaptive choice of τ and a global choice of h. However, in some cases this procedure presents some drawbacks which we briefly describe below.

II.2 Article 2 : Non parametric adaptive estimation of conditional probabilities of rares events and extreme quantiles with simultaneous determination of the threshold and the bandwidth

Our simulations show that the global choice is less efficient than the choice based on an adaptive local procedure, especially when the number of observations in the time series is very large, since the computational time increases with the size of the sample. A simple remedy to this is to split the series into smaller parts and then to apply the same approach for partitioned series. The second drawback is that the choice of h based on the minimization of the cross validation loss (I.35) is sensible to the choice of the quantile level p. By numerical computations, we observed that the precision may become worse when the values of p are too close to 1, for instance when $p = 0.9999$. A remedy for this is to choose h_n by cross validation for p' of smaller order, and to apply the selected bandwidth for estimation quantiles of higher order $p > p'$. For example, the selection could be performed by cross validation with $p' = 0.99$ and the selected bandwidth, say h', used to estimate quantiles of order $p = 0.9999$. Finally, a common drawback of the procedures based on a global choice of the bandwidth is the lack of adaptability against the change of the regularity of the estimated function, which is also the case in the present setting.

For the reasons described above, we turned to a local procedure which seems to have some advantages if the estimated function is not of homogenous regularity on the estimated interval. It is the aim of this section is to give a local in the time t selection procedure for the simultaneous choice the bandwidth h and threshold τ. For this we shall combine the two procedures stated in Sections 2.4.1 and 2.2.2 : the first one chooses the threshold τ while the bandwidth h is fixed and the second one chooses the bandwidth h while the threshold τ is fixed. The procedure for simultaneous choice of the threshold τ and the bandwidth h looks much like that for the adaptive choice of h, but is used in conjunction with the selection procedure for choosing the threshold τ.

As in the case of adaptive choice of bandwidth h in Section 2.2.2, let $t \in [0, T_{\max}]$ and M be a positive integer. Consider an increasing sequence of bandwidths $0 < \hbar_1 < \hbar_2 < ... < \hbar_M$. As

before, we shall use in the sequel the uniform grid $\hbar_m = \hbar_1 + (m-1)\delta$, where $\hbar_1$ and δ are positive numbers, to be calibrated. . The sequence $\{\mathbf{h_m}\}$ defines a family $\mathcal{I}_t = \{I_{t,\hbar_m} : m = 1, ..., M\}$ of nested intervals of the form $I_{t,\hbar_m} = [t - \hbar_m, t + \hbar_m] \cap [0, T_{\max}]$.

Similarly, for the choice of the threshold τ we need a sequence of thresholds which we proceed to introduce. Assume that the bandwidth $h = h_m$ is given. Let $Y_1 \geq ... \geq Y_{n_{t,h}}$ be the order statistics pertaining to the observations $\{X_{t_i} : t_i \in I_{t,h}\}$, where $n_{t,h}$ is the cardinal of this set. We will make use of the following two test statistics which have been used in Section 2.4.1 . The first one is

$$LR_{t,h}(s,\tau) = \widehat{n}_{t,h,s,\tau}\mathcal{K}\left(\widehat{\mu}_{t,h,s,\tau}, \widehat{\theta}_{t,h,s}\right) + \widehat{n}_{t,h,\tau}\mathcal{K}\left(\widehat{\theta}_{t,h,\tau}, \widehat{\theta}_{t,h,s}\right). \tag{II.32}$$

where $\widehat{\theta}_{t,h,\tau}$ is the maximum likelihood estimator of θ given by (II.5),

$$\widehat{\mu}_{t,h,s,t} = \frac{\widehat{n}_{t,h,s}}{\widehat{n}_{t,h,s,\tau}}\widehat{\theta}_{t,h,s} - \frac{\widehat{n}_{t,h,\tau}}{\widehat{n}_{t,h,s,\tau}}\widehat{\theta}_{t,h,\tau},$$

and

$$\widehat{n}_{t,h,s,\tau} = \sum_{t_i \in I_{t,h}} 1_{\{s < X_{t_i} \leq \tau\}}.$$

The second one is the penalized quasi-log-likelihood

$$\mathcal{L}^{\text{Pen}}_{t,h}(s,\tau) = \widehat{n}_{t,h,\tau}\mathcal{K}\left(\widehat{\theta}_{t,h,\tau}, \widehat{\theta}_{t,h,s}\right). \tag{II.33}$$

In the following, D and z denote two critical values of the testing procedures obtained by simulations. The algorithm for the simultaneous choice of h and τ reads as follows :

Step 1. Set $m = m_0$.

Step 2.1. Set $k = k_0$ and $h = h_m$.

Step 2.2. Compute the test statistic

$$\mathbf{Z}_h(Y_k) = \max_{\delta'\widehat{n}_{t,h,Y_k} \leq \widehat{n}_{t,h,Y_l} \leq (1-\delta'')\widehat{n}_{t,h,Y_k}} LR_{t,h}(Y_k, Y_l). \tag{II.34}$$

Step 2.3. If for $k \leq M_{t,h}$ we have $\mathbf{Z}_h(Y_k) > D$, then we let $k^* = k-1$ and define the adaptive index $\widehat{k}$ by maximizing the penalized likelihood

$$\widehat{k} = \underset{\eta'\widehat{n}_{t,h,Y_{k^*}} \leq \widehat{n}_{t,h,Y_l} \leq (1-\eta'')\widehat{n}_{t,h,Y_{k^*}}}{\arg\max} \mathcal{L}^{\mathrm{Pen}}_{t,h}(Y_{k^*}, Y_l), \tag{II.35}$$

where $\mathcal{L}^{\mathrm{Pen}}_{t,h}$ is defined by (II.33). If for $k \leq M_{t,h}$ we have $\mathbf{Z}_h(Y_k) \leq D$ then : if $k < M_{t,h}$ we increase k by 1 and return to Step 2.2 ; if $k = M_{t,h}$ set the adaptive index as $\widehat{k} = M_{t,h}$. We define the adaptive threshold as $\widehat{\tau}_m = Y_{\widehat{k}}$ and go to Step 3.

Step 3. Compute the test statistic (II.24) for testing the homogeneity on I_{t,h_m} :

$$\mathbb{Z}_{\hbar_m}(\widehat{\tau}_m) = \max_{\eta'\widehat{n}_{t,\hbar_m,\widehat{\tau}_m} \leq \widehat{n}_{t,\hbar_l,\widehat{\tau}_m} \leq (1-\eta'')\widehat{n}_{t,\hbar_m,\widehat{\tau}_m}} LR_{t,\widehat{\tau}_m}(\hbar_m, \hbar_l).$$

Step 4. If for $m \leq M$ a deviation from homogeneity is detected, i.e. $\mathbb{Z}_{\hbar_m}(\widehat{\tau}_m) > z$ let $m^* = m-1$ and define the adaptive index by maximizing the penalized likelihood

$$\widehat{m} = \underset{\eta'\widehat{n}_{t,\hbar_{m^*},\widehat{\tau}_{m^*}} \leq \widehat{n}_{t,\hbar_l,\widehat{\tau}_{m^*}} \leq (1-\eta'')\widehat{n}_{t,\hbar_{m^*},\widehat{\tau}_{m^*}}}{\arg\max} LR^{\mathrm{Pen}}_{t,\widehat{\tau}_{m^*}}(\hbar_{m^*}, \hbar_l), \tag{II.36}$$

where $LR^{\mathrm{Pen}}_{t,\tau}$ is defined in (II.25). If for $m \leq M$ the test statistic computed above does not detect a deviation from the homogeneity, i.e. $\mathbb{Z}_{\hbar_m}(\widehat{\tau}_m) \leq z$, then : if $m < M$ increase m by 1 and return to Step 2 ; if $m = M$ set the adaptive index as $\widehat{m} = M$.

Step 5. We end the algorithm by defining the adaptive bandwidth, threshold and estimator by $\widehat{h}_t = \hbar_{\widehat{m}}$, $\widehat{\tau}_t = \widehat{\tau}_{\widehat{m}}$ and $\widehat{\theta}_{t,\widehat{h}_t,\widehat{\tau}_t}$.

The choice of the parameters for this procedure is discussed in Section 2, Chaper III.

2.4 Proofs

Denote for brevity $\mathcal{H} = \{\hbar_m : k = 1, ..., M\}$. For any $t \in [0, T_{\max}]$, $\tau \geq x_0$, $\theta > 0$ and $h > 0$, define

$$d_{t,\tau,h,\theta} = \sum_{t_i \in I_{t,h}} \chi^2(F_{t_i}, F_{t_i,\tau,\theta}).$$

Proposition II.10. *Let $t \in [0, T_{\max}]$. For any $\tau \geq x_0$, $\theta > 0$, $h > \hbar_{k_0}$ and any $y > 0$, we have*

$$\mathbb{P}\left(\max_{\hbar_l \in \mathcal{H}, \hbar_l < h} LR_{t,\tau}(h, \hbar_l) \leq 2y + 4\log\left(\frac{n_{t,h}^2}{2}\right) + d_{t,\tau,h,\theta}\right) \geq 1 - 2\exp\left(-\frac{y}{2}\right), \tag{II.37}$$

where $n_{t,h}$ is the number of t_i in the interval $I_{t,h}$.

Proof. Let $y > 0$ be fixed . Since for any $\hbar_l \in \mathcal{H}$ such that $\hbar_l < h$,

$$\begin{aligned} LR_{t,\tau}(h, \hbar_l) &= \widehat{n}_{t,\hbar_l,\tau}\mathcal{K}(\widehat{\theta}_{t,\hbar_l,\tau}, \widehat{\theta}_{t,h,\tau}) + \widehat{n}_{t,h,\hbar_l,\tau}\mathcal{K}(\widehat{\theta}_{t,h,\hbar_l,\tau}, \widehat{\theta}_{t,h,\tau}) \\ &= \widehat{n}_{t,\hbar_l,\tau}\mathcal{K}(\widehat{\theta}_{t,\hbar_l,\tau}, \theta) + \widehat{n}_{t,h,\hbar_l,\tau}\mathcal{K}(\widehat{\theta}_{t,h,\hbar_l,\tau}, \theta) - \widehat{n}_{t,h,\tau}\mathcal{K}(\widehat{\theta}_{t,h,\tau}, \theta). \end{aligned}$$

we have

$$LR_{t,\tau}(h, \hbar_l) \leq \widehat{n}_{t,\hbar_l,\tau}\mathcal{K}(\widehat{\theta}_{t,\hbar_l,\tau}, \theta) + \widehat{n}_{t,h,\hbar_l,\tau}\mathcal{K}(\widehat{\theta}_{t,h,\hbar_l,\tau}, \theta).$$

Denote

$$\begin{aligned} \Omega_t(h) &= \cap_{\hbar_l \in \mathcal{H}, \hbar_l < h}\Omega_{t,\hbar_l}, \text{ where} \\ \Omega_{t,\hbar_l} &= \left\{\widehat{n}_{t,\hbar_l,\tau}\mathcal{K}(\widehat{\theta}_{t,\hbar_l,\tau}, \theta) \leq y_l^{(1)}, \widehat{n}_{t,h,\hbar_l,\tau}\mathcal{K}(\widehat{\theta}_{t,h,\hbar_l,\tau}, \theta) \leq y_l^{(2)}\right\} \end{aligned}$$

with $y_l^{(1)}, y_l^{(2)} > 0$. According to Proposition I.13 in Chapter I, we have for any $\hbar_l \in \mathcal{H}, \hbar_l < h$

$$\begin{aligned} \mathbb{P}(\bar{\Omega}_{t,\hbar_l}) &\leq n_{t,\hbar_l}\exp\left(-\frac{y_l^{(1)}}{2} + \frac{1}{2}\sum_{t_i \in I_{t,\hbar_l}}\chi^2(F_{t_i}, F_{t_i,\tau,\theta})\right) \\ &+ (n_{t,h} - n_{t,\hbar_l})\exp\left(-\frac{y_l^{(2)}}{2} + \frac{1}{2}\sum_{t_i \in I_{t,h}\setminus I_{t,\hbar_l}}\chi^2(F_{t_i}, F_{t_i,\tau,\theta})\right). \end{aligned}$$

Therefore, by taking

$$\begin{aligned} y_l^{(1)} &= y + 2\log n_{t,h} + 2\log n_{t,\hbar_l} + \sum_{t_i \in I_{t,\hbar_l}}\chi^2(F_{t_i}, F_{t_i,\tau,\theta}) \\ y_l^{(2)} &= y + 2\log n_{t,h} + 2\log(n_{t,h} - n_{t,\hbar_l}) + \sum_{t_i \in I_{t,h}\setminus I_{t,\hbar_l}}\chi^2(F_{t_i}, F_{t_i,\tau,\theta}), \end{aligned}$$

we deduce,

$$\mathbb{P}(\bar{\Omega}_{t,\hbar_l}) \leq 2\exp\left(-\frac{y}{2} - \log n_{t,h}\right).$$

Hence

$$\begin{aligned}\mathbb{P}(\Omega_t(h)) &\geq 1 - \sum_{\hbar_l \in \mathcal{H}, \hbar_l \leq h} \mathbb{P}(\bar{\Omega}_{t,\hbar_l}) \\ &\geq \sum_{\hbar_l \in \mathcal{H}, \hbar_l \leq h} 2\log\left(-\frac{y}{2} - \log n_{t,h}\right) \\ &\geq 1 - 2n_{t,h}\log\left(-\frac{y}{2} - \log n_{t,h}\right) \\ &\geq 1 - 2\exp\left(-\frac{y}{2}\right).\end{aligned}$$

On the set $\Omega_t(h)$, for any $\hbar_l \in \mathcal{H}, \hbar_l < h$

$$\begin{aligned}LR_{t,\tau}(h,\hbar_l) &\leq y_l^{(1)} + y_l^{(2)} \\ &= 2y + 4\log n_{t,h} + 2\log\left(n_{t,\hbar_l}(n_{t,h} - n_{t,\hbar_l})\right) + \sum_{t_i \in I_{t,h}} \chi^2(F_{t_i}, F_{t_i,\tau,\theta}) \\ &\leq 2y + 4\log\left(\frac{n_{t,h}^2}{2}\right) + d_{t,\tau,h,\theta}.\end{aligned}$$

It follows that on $\Omega_t(h)$

$$\max_{\hbar_l \in \mathcal{H}, \hbar_l < h} LR_{t,\tau}(h,\hbar_l) \leq 2y + 4\log\left(\frac{n_{t,h}^2}{2}\right) + d_{t,\tau,h,\theta}$$

which completes the proof. □

2.4.1 Proof of Proposition II.4

Let $t \in [0, T_{\max}]$. Using the Proposition II.10 with $y = 3\log n_{t,h_n}$, we have

$$1 - \frac{2}{n_{t,h_n}^{3/2}} \leq \mathbb{P}\left(\max_{\hbar_l \in \mathcal{H}, \hbar_l < h} LR_{t,\tau_n}(h,\hbar_l) \leq 6\log n_{t,h_n} + 4\log\left(\frac{n_{t,h}^2}{2}\right) + d_{t,\tau_n,h,\theta_n}\right). \tag{II.38}$$

From the condition C1, there exist a constant $c > 0$ such that

$$d_{t,\tau_n,h_n,\theta_n} \leq c\log n_{t,h_n}.$$

Hence,

$$\begin{aligned} 6\log n_{t,h_n} + 4\log\frac{n_{t,h}^2}{2} + d_{t,\tau_n,h,\theta_n} &\leq 6\log n_{t,h_n} + 4\log\frac{n_{t,h_n}^2}{2} + d_{t,\tau_n,h_n,\theta_n} \\ &\leq (14+c)\log n_{t,h_n}. \end{aligned}$$

Therefore, from (II.38) and the fact that

$$\mathbb{Z}_h(\tau_n) \leq \max_{\hbar_l\in\mathcal{H},\hbar_l<h} LR_{t,\tau_n}(h,\hbar_l),$$

we obtain

$$\begin{aligned} 1-\frac{2}{n_{t,h_n}^{3/2}} &\leq \mathbb{P}\left(\max_{\hbar_l\in\mathcal{H},\hbar_l<h} LR_{t,\tau_n}(h,\hbar_l) \leq 6\log n_{t,h_n} + 4\log\left(\frac{n_{t,h}^2}{2}\right) + d_{t,\tau_n,h,\theta_n}\right) \\ &\leq \mathbb{P}\left(\max_{\hbar_l\in\mathcal{H},\hbar_l<h} LR_{t,\tau_n}(h,\hbar_l) \leq (14+c)\log n_{t,h_n}\right) \\ &\leq \mathbb{P}\left(\mathbb{Z}_h(\tau_n) \leq (14+c)\log n_{t,h_n}\right). \end{aligned}$$

The result follows by taking $c^* = 14 + c$.

2.4.2 Proof of Theorem II.5

For sake of brevity, we denote $\widehat{n}_k = \widehat{n}_{t,\hbar_k,\tau_n}$, $\widehat{\theta}_k = \widehat{\theta}_{t,\hbar_k,\tau_n}$.

With c^* determined in Proposition II.4, we denote

$$\Omega_t^*(h_n) = \bigcap_{\hbar_r\in\mathcal{H},\hbar_r\leq h_n} \left\{\mathbb{Z}_{\hbar_r}(\tau_n) \leq c^*\log n_{t,h_n}\right\}. \tag{II.39}$$

According to that Proposition, we obtain

$$\begin{aligned} \mathbb{P}\left(\Omega_t^*(h_n)\right) &\geq 1 - \sum_{\hbar_r\in\mathcal{H},\hbar_r\leq h_n} \left\{\mathbb{Z}_{\hbar_r}(\tau_n) > c^*\log n_{t,h_n}\right\} \\ &\geq 1 - \sum_{\hbar_r\in\mathcal{H},\hbar_r\leq h_n} \frac{2}{n_{t,h_n}^{3/2}} \geq 1 - \frac{2}{n_{t,h_n}^{1/2}}. \end{aligned}$$

In the following, we assume that $k_{step} = 1$.

We first compare $\widehat{\theta}_{t,h_n,\tau_n}$ with $\widehat{\theta}_{m^*}$. From the definition of m^* and the Proposition II.4, it

holds $h_n < \hbar_{m^*+1}$. Hence $h_n \leq \hbar_{m^*}$ and $\widehat{n}_{t,h_n,\tau_n} \leq \widehat{n}_{m^*}$. We define the sequence of positive integer numbers $m_i, i = 0, ..., i^*$ such that $m_0 = m^*$ and $\eta'\widehat{n}_{m_{i-1}} \leq \widehat{n}_{m_i} \leq \frac{1}{2}\widehat{n}_{m_{i-1}} \leq (1-\eta'')\widehat{n}_{m_{i-1}}$, for $i = 1, ..., i^*$, where i^* such that $\eta'\widehat{n}_{m_{i^*}} \leq \widehat{n}_{\tau_n}^{(I_{t,h_n})} \leq \frac{1}{2}\widehat{n}_{m_{i^*}} \leq (1-\eta'')\widehat{n}_{m_{i^*}}$. Denote $\widehat{n}_{m_{i^*+1}} = \widehat{n}_{t,h_n,\tau_n}$ and $\widehat{\theta}_{m_{i^*+1}} = \widehat{\theta}_{t,h_n,\tau_n}$. It is obvious that on $\Omega_t^*(h_n)$, we have

$$\mathbb{Z}_{h_n}(\tau_n) \leq c^* \log n_{t,h_n}$$

and

$$\mathbb{Z}_{\hbar_r}(\tau_n) \leq c^* \log n_{t,h_n}, \text{ for } k_0 \leq r \leq \widehat{k} - 1.$$

Therefore, by (II.23), we obtain

$$\widehat{n}_{m_{i+1}} \mathcal{K}(\widehat{\theta}_{m_{i+1}}, \widehat{\theta}_{m_i}) \leq z = c^* \log n_{t,h_n}, i = 0, ..., i^*.$$

This implies

$$\sum_{i=0}^{i^*} \sqrt{\mathcal{K}(\widehat{\theta}_{m_{i+1}}, \widehat{\theta}_{m_i})} \leq z^{1/2} \sum_{i=0}^{i^*} \widehat{n}_{m_{i+1}}^{-1/2}.$$

Taking into account that $\widehat{n}_{m_{i+1}} \leq \frac{1}{2}\widehat{n}_{m_i}$, for $i = 0, ..., i^*$, we have

$$\sum_{i=0}^{i^*} \widehat{n}_{m_{i+1}}^{-1/2} \leq \widehat{n}_{m_{i^*+1}}^{-1/2} \sum_{i=0}^{i^*} 2^{-(i^*-i)/2} \leq (2+\sqrt{2})\widehat{n}_{m_{i^*+1}}^{-1/2},$$

According to Lemma 8.1 and 8.2 in [15], for n sufficiently large

$$\sqrt{\mathcal{K}(\widehat{\theta}_{m_{i^*+1}}, \widehat{\theta}_{m_0})} \leq \frac{3}{2} \sum_{i=0}^{i^*} \sqrt{\mathcal{K}(\widehat{\theta}_{m_{i+1}}, \widehat{\theta}_{m_i})} \leq \frac{3}{2}(2+\sqrt{2}) z^{1/2} \widehat{n}_{m_{i^*+1}}^{-1/2},$$

and

$$\sqrt{\mathcal{K}(\widehat{\theta}_{\widehat{k}-1}, \widehat{\theta}_{t,h_n,\tau_n})} = \sqrt{\mathcal{K}(\widehat{\theta}_{m_0}, \widehat{\theta}_{m_{i^*+1}})}$$

$$\leq \frac{9}{4}(2+\sqrt{2}) z^{1/2} \widehat{n}_{m_{i^*+1}}^{-1/2}$$

$$= \frac{9}{4}(2+\sqrt{2})\sqrt{\frac{z}{\widehat{n}_{m_{i^*+1}}}} \tag{II.40}$$

We now compare $\widehat{\theta}_{\widehat{m}}$ with $\widehat{\theta}_{m^*}$. By the definition of m^*, we have

$$\mathbb{Z}_{h_{m^*}}(\tau_n) \leq z.$$

It follows that

$$LR^{Pen}_{t,\tau_n}(\hbar_{m^*}, \hbar_{\widehat{m}}) = \widehat{n}_{\widehat{m}} \mathcal{K}\left(\widehat{\theta}_{\widehat{m}}, \widehat{\theta}_{m^*}\right) \leq z.$$

Therefore,

$$\sqrt{\mathcal{K}\left(\widehat{\theta}_{\widehat{m}}, \widehat{\theta}_{m^*}\right)} \leq z^{1/2}\widehat{n}_{\widehat{m}}^{-1/2} \leq \sqrt{\frac{z}{\eta'\widehat{n}_{m^*}}} \leq \sqrt{\frac{z}{\eta'\widehat{n}_{m_{i^*+1}}}}. \tag{II.41}$$

Combining (II.40) and (II.41), by Lemma 8.1 and 8.2 in[15], it follows that, on the set $\Omega^*_t(h_n)$, we have

$$\sqrt{\mathcal{K}(\widehat{\theta}_{\widehat{m}}, \widehat{\theta}_{t,h_n,\tau_n})} \leq \sqrt{c\frac{z}{\widehat{n}_{m_{i^*+1}}}} = \sqrt{cc^*\frac{\log n_{t,h_n}}{\widehat{n}_{t,h_n,\tau_n}}}.$$

with some positive constant c. Taking into account (II.39), we obtain

$$\mathbb{P}\left(\mathcal{K}(\widehat{\theta}_{\widehat{m}}, \widehat{\theta}_{t,h_n,\tau_n}) \leq cc^*\frac{\log n_{t,h_n}}{\widehat{n}_{t,h_n,\tau_n}}\right) \to 1 \text{ as } n \to \infty.$$

Using Lemma 8.5 in Durrieu et al (2014)[56] and $n_{t,h_n} \leq n$, we deduce that

$$\mathbb{P}\left(\mathcal{K}(\widehat{\theta}_{\widehat{m}}, \widehat{\theta}_{t,h_n,\tau_n}) \leq cc^*\frac{\log n_{t,h_n}}{\bar{n}_{t,h_n,\tau_n}}\right) \to 1 \text{ as } n \to \infty,$$

and

$$\mathbb{P}\left(\mathcal{K}(\widehat{\theta}_{\widehat{m}}, \widehat{\theta}_{t,h_n,\tau_n}) \leq cc^*\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right) \to 1 \text{ as } n \to \infty,$$

which complete the proof.

Chapitre III

Simulation results

Pour une mise en oeuvre de la méthode statistique développée dans les Chapitres I et II, une implémentation en R a été réalisée. Nous allons dans ce Chapitre présenter les modèles simulés ainsi qu'une présentation et une discussion des résultats obtenus. Nous avons effectué ces simulations numériques afin :

1. d'étudier le comportement et la sensibilité des méthodes étudiées sur trois modèles (modèle de mélange, modèle de Burr et modèle de Burr dépendant),
2. de comparer les résultats pour 2 méthodes d'estimation de la taille de fenêtre h (méthodes de validation croisée et adaptative) avec une méthode optimale,
3. d'évaluer la qualité des estimateurs des quantiles extrêmes proposés par rapport à la taille de échantillon n,
4. d'étudier le comportement de la procédure sur des données dépendantes.

Nous commençons par choisir les paramètres de l'algorithme. Nous nous intéressons dans ces simulations à des modèles de régression avec design fixe.

Nous supposons que K est un noyau symétrique à support compact sur $[-1, 1]$. Nous précisons les noyaux utilisés dans nos simulations dans le tableau III.1. Pour chaque $t \in [0, T_{\max}]$, nous utilisons par la suite une famille de poids $\{W_{t,h}(t_i) = K(\frac{t-t_i}{h}) : t - t_i \in [0, T_{\max}]\}$, où $t_i = i/n\, T_{\max} \in [0, T_{\max}], i = 1, \ldots, n$.

Tableau III.1 – Quelques fonctions noyaux

Nom	$K(u)$
Noyau rectangulaire	$K(u) = 1_{[-1,1]}(u)$
Noyau triangulaire	$K(u) = (1 - \|u\|)\, 1_{[-1,1]}(u)$
Noyau d'Epanechnikov	$K(u) = \frac{3}{4}(1 - u^2)\, 1_{[-1,1]}(u)$
Noyau biweight	$K(u) = \frac{15}{16}(1 - u^2)^2\, 1_{[-1,1]}(u)$
Noyau Gaussien tronqué	$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) 1_{[-1,1]}(u)$

1 Procédure de sélection du seuil τ : choix des paramètres

Nous rappelons les choix des paramètres donnés dans la partie Simulation 2.5 du Chapitre I. Nous fixons $\delta_0 = \frac{1}{10}$, $\delta' = \frac{1}{4}$, $\delta'' = \frac{1}{20}$ et $M_{grid} = 100$. Nous choisissons le noyau gaussien tronqué symétrique et à support compact sur $[-1, 1]$ mais il est possible de choisir d'autres noyaux comme par exemple les noyaux rectangulaire, triangulaire, Epanechnikov et biweight définis dans la Table III.1.

Pour déterminer la valeur critique notée D, nous utilisons la statistique de test $\mathbf{Z}_h(Y_k)$ décrite par (II.34) dans la procédure de selection du seuil τ du paragraphe 2.4 sous l'hypothèse nulle qui stipule que $X_{t_i}, X_{t_2}, \ldots, X_{t_n}$ sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de loi de Pareto $G_{\tau,\theta}$ avec $\tau, \theta > 0$. Puisque sous l'hypothèse nulle la distribution de $\mathbf{Z}_h(Y_k)$ ne dépend pas de τ et θ, on peut choisir la loi de Pareto standard $G_{1,1}$.

Nous simulons donc $N_{MC} = 2000$ échantillons $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ i.i.d. de distribution de Pareto standard de tailles $n = 1000, 2500, 5000$ et 10000. Pour chaque échantillon, nous calculons la statistique de test

$$T_n = \sup_{k_0 \le k = k_0 + i k_{step} \le n} \mathbf{Z}_h(Y_k)$$

au point $t = 0.5$. Les Figures III.1, III.2, III.3, III.4 et III.5 représentent la distribution empirique de la statistique de test T_n associée aux différents noyaux pour $2\,n\,h \in \{200, 500, 1000, 2000\}$.

Tableau III.2 – Valeurs critiques associées aux noyaux

Noyaux	D
Rectangulaire	10.0
Triangulaire	6.9
Epanechnikov	6.1
Biweight	12.3
Gaussien tronqué	3.6

Pour chaque noyau utilisé et pour chaque valeur de $2nh$, la valeur critique est déterminée comme étant le 0.99-quantile empirique de la statistique T_n. Les valeurs critiques pour les fonctions noyaux du Tableau III.1 sont données dans le Tableau III.2. Ces Figures montrent que les valeurs des 0.99-quantiles empiriques sont voisines par rapport au changement de $2nh$.

La Figure III.6 présente un exemple de choix du seuil de rupture $\widehat{s}$ et du seuil adaptatif $\widehat{\tau}_n$ en fonction de la valeur critique D. Nous observons que le seuil adaptatif choisi $\widehat{\tau}_n$ est stable par rapport au changement de la valeur critique D.

2 Procédure de sélection simultanée du seuil τ et de la taille de la fenêtre h : choix des paramètres

Dans la procédure décrite dans le Chapitre II, nous devons déterminer deux valeurs critiques nommées D et z. Le choix des paramètres $\delta_0, \delta', \delta''$ et de la première valeur critique D sont présentés dans le paragraphe précédent. Nous choisissons $\delta_0 = 0.1, \delta' = 0.25, \delta'' = 0.25$ et $D = 10.0$ correspondant au noyau rectangulaire. Nous fixons aussi $\eta' = 0.2$, $\eta'' = 0.05$ et une suite de taille de fenêtre $\hbar_m = \hbar_1 + (m-1)\delta$, $m = 1, \ldots, M$ où $\hbar_1 = 0.025, \delta = 10^{-3}, 5.10^{-4}$, $\hbar_M = 0.5$ et $M = 476$.

Pour obtenir la seconde valeur critique z, nous utilisons la statistique de test $\mathbb{Z}_{\hbar_m}(\widehat{\tau}_m)$ (donnée par II.24) sous l'hypothèse nulle qui stipule que $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de loi de Pareto $G_{\tau,\theta}$ avec $\tau, \theta > 0$.

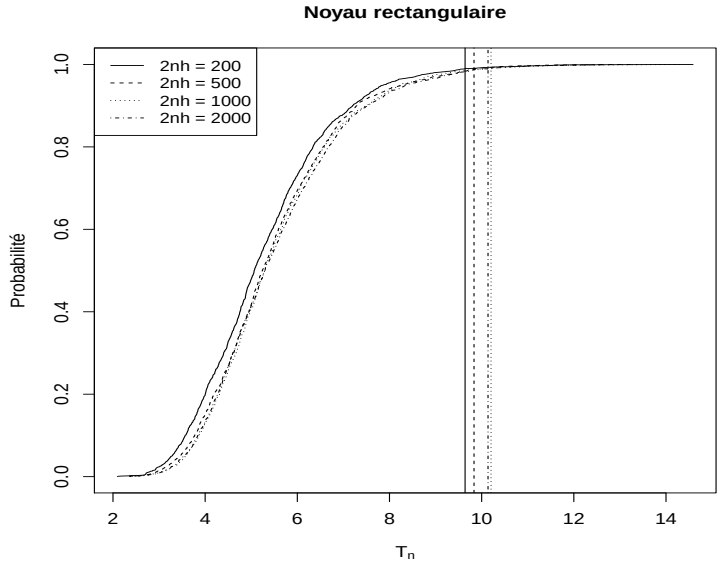

Figure III.1 – Distribution empirique de la statistique de test T_n avec le noyau rectangulaire pour $2nh = 200, 500, 1000, 2000$. Les lignes verticales correspondent aux quantiles empiriques d'ordre 0.99.

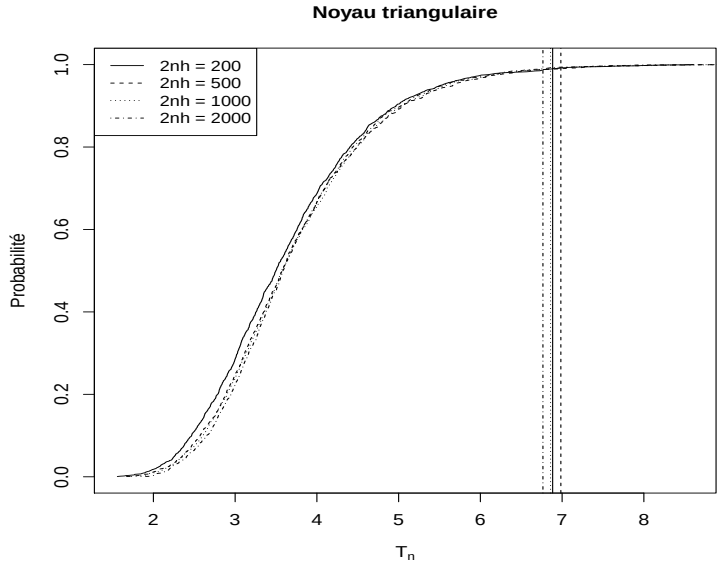

Figure III.2 – Distribution empirique de la statistique de test T_n avec le noyau triangulaire pour $2nh = 200, 500, 1000, 2000$. Les lignes verticales correspondent aux quantiles empiriques d'ordre 0.99.

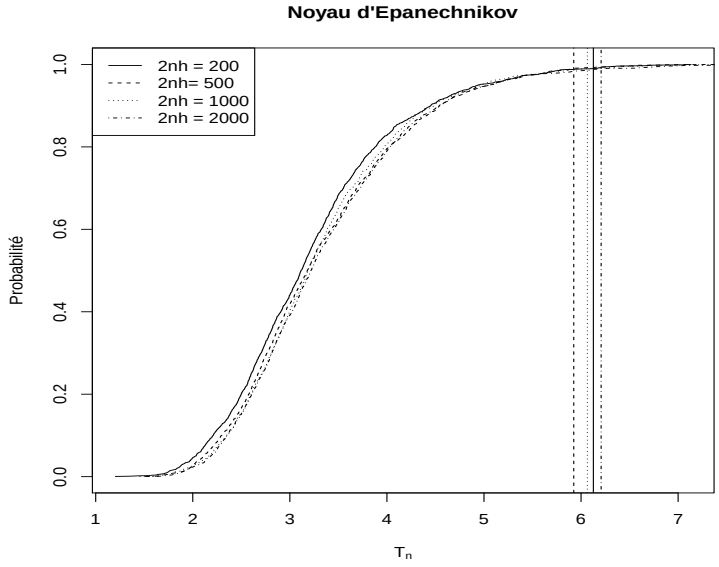

Figure III.3 – Distribution empirique de la statistique de test T_n avec le noyau d'Epanechnikov pour $2nh = 200, 500, 1000, 2000$. Les lignes verticales correspondent aux quantiles empiriques d'ordre 0.99.

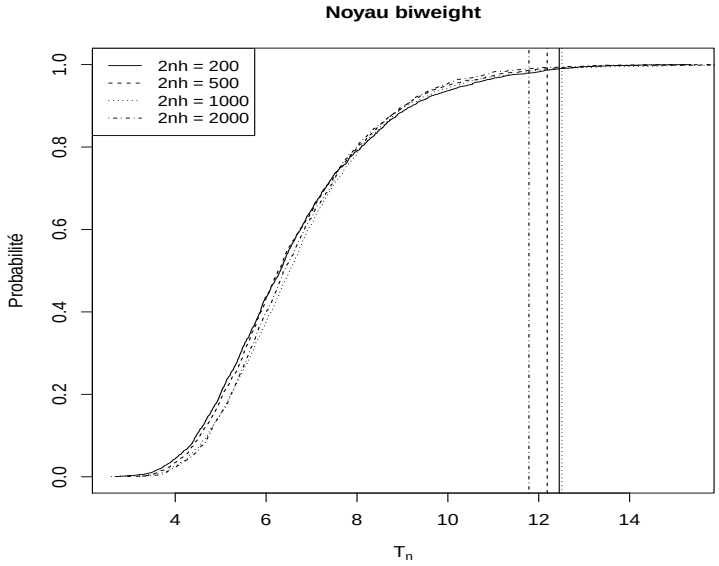

Figure III.4 – Distribution empirique de la statistique de test T_n avec le noyau biweight pour $2nh = 200, 500, 1000, 2000$. Les lignes verticales correspondent aux quantiles empiriques d'ordre 0.99.

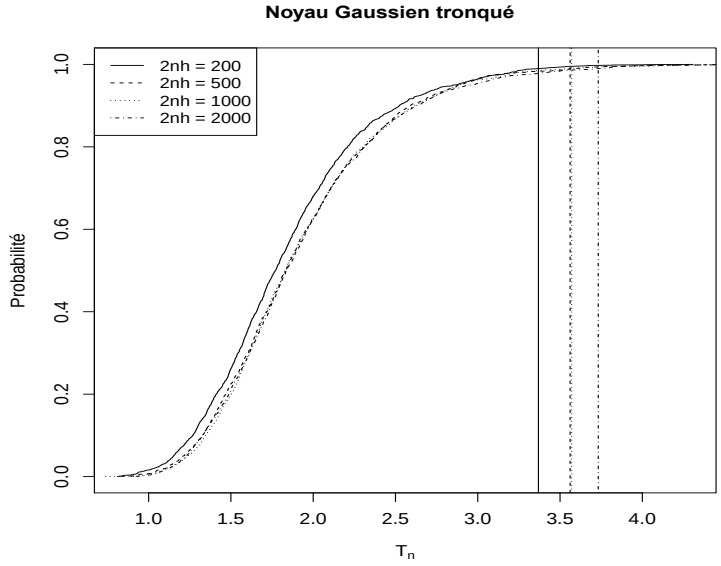

Figure III.5 – Distribution empirique de la statistique de test T_n avec le noyau Gaussien tronqué pour $2nh = 200, 500, 1000, 2000$. Les lignes verticales correspondent aux quantiles empiriques d'ordre 0.99.

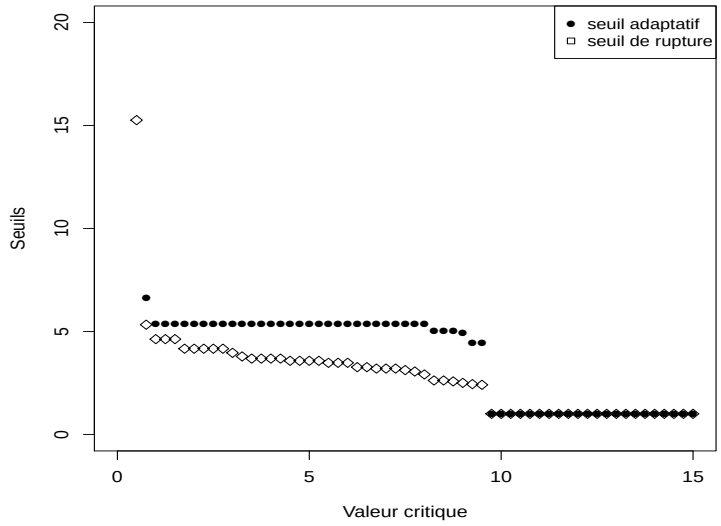

Figure III.6 – Choix du seuil de rupture et du seuil adaptatif en fonction de la valeur critique D.

Tableau III.3 – Valeurs critiques z

$2n\hbar_m$	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
z	2.88	3.85	4.08	4.13	4.36	4.21	4.22	4.38	4.43	4.52

A nouveau, sous l'hypothèse nulle la distribution de $\mathbb{Z}_{\hbar_m}(\widehat{\tau}_m)$ ne dépend pas de τ et θ. On peut donc choisir la loi de Pareto standard $G_{1,1}$ pour determiner z.

Nous simulons $N_{MC} = 2000$ échantillons $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ i.i.d. de distribution de Pareto standard de tailles $n = 5000$ et $n = 50000$. Pour chaque échantillon et chaque $\hbar_m$, nous calculons la statistique de test

$$z_m = \mathbb{Z}_{\hbar_m}(\widehat{\tau}_m)$$

au point $t = 0.5$, où $\hbar_m \geq \hbar_{m_0} = 0.05$ et $\widehat{\tau}_m = \min\{X_{t_i}, t_i \in [t - \hbar_m, t + \hbar_m]\}$. Pour chaque valeur de $\hbar_m$, la valeur critique z est déterminée comme étant le 0.95-quantile empirique de la statistique z_m. Le Tableau III.3 représente des valeurs critiques en fonction de $2n\hbar_m$ avec $n = 5000$.

La Figure III.7 présente un exemple de choix de la fenêtre de rupture $\hbar_{m^*}$ et la taille de la fenêtre adaptative $\widehat{h}_t$ en fonction de la valeur critique z. Nous observons que la taille de la fenêtre adaptative choisie $\widehat{h}_t$ reste stable par rapport au changement de la valeur critique z. C'est pourquoi dans les simulations, il est possible de choisir comme valeur critique une valeur dans l'intervalle $[2.88, 4.52]$. Nous fixons $z = 4.10$ qui corresponds à la valeur mediane.

3 Modèle de mélange

Nous considérons le modèle de mélange décrit dans le paragraphe 2.5 du Chapitre I :

$$F_t(x) = C(1 - x^{-1/\theta_t}) + (1 - C)\left(1 - x^{-1/\theta_t - 5}\right), \; x \geq 1, 0 \leq t \leq 1, \tag{III.1}$$

où $C = 0.75$ et $\theta_t = 0.5 + 0.25 \sin(2\pi t)$.

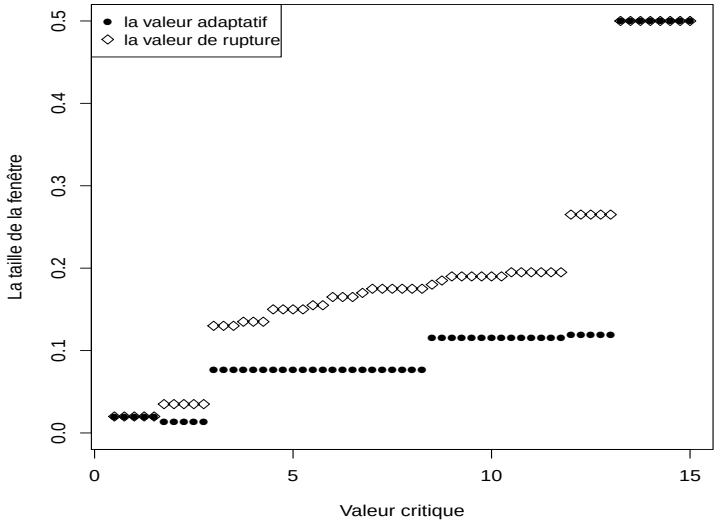

Figure III.7 – Choix de la taille de la fenêtre de rupture et la taille de la fenêtre adaptative en fonction de la valeur critique z.

Nous simulons à partir du modèle (III.1) $N_{MC} = 1000$ échantillons $X_{t_1}, \ldots, X_{t_n}$ de tailles $n = 5000$ et $n = 50000$ avec $t_i = i/n$, $i = 1, \ldots, n$. Notons $T_{grid} = \{0.01, 0.02, \ldots, 0.99\}$ un ensemble de points d'une grille sur intervalle $[0, 1]$. Pour chaque échantillon simulé, nous calculons les estimateurs du p-quantile $\widehat{q}_p(t_i)$ pour $p \in \{0.99, 0.999, 0.9999\}$, $t_i \in T_{grid}$ avec τ et h estimés par les procédures proposées.

Puisque dans les simulations les quantiles théoriques $F_t^{-1}(p)$ sont connus, nous pouvons déterminer une taille de fenêtre h "optimale" notée dans la suite h_{opt} de la façon suivante. Considérons $\mathcal{H} = \{h_m : h_m = h_0 q^m, m = 1, \ldots, M_h\}$ avec $q > 1$, $h_0 > 0$ et M_h grand. Pour chaque h_m, l'erreur relative intégrée notée ISRE (Integrated Squared Relative Error) entre le quantile estimé et le vrai quantile est défini par

$$ISRE(h_m, p) = \frac{1}{card(T_{grid})} \sum_{t_i \in T_{grid}} \phi\left(\widehat{q}_p(t_i, h_m), F_{t_i}^{-1}(p)\right), \tag{III.2}$$

où T_{grid} est une suite de points d'une grille régulière sur $[0, T_{\max}]$, $\widehat{q}_p(t_i, h_m)$ désigne l'estimateur

du quantile d'ordre p donnée par (I.4), et $\phi(x,y) = (\log x - \log y)^2$, $x, y > 0$. La valeur h_{opt} s'obtient par minimisation par rapport à h_m de $ISRE$ pour p fixé.

Pour évaluer et comparer la performance des procédures proposées, nous calculons l'$ISRE$ de l'estimateur des quantiles :

– avec h_{opt}

$$ISRE_{opt}^{(j)} = \frac{1}{card(T_{grid})} \sum_{t_i \in T_{grid}} \log^2 \left(\widehat{q}_p^{(j)}(t_i, h_{opt}) / F_{t_i}^{-1}(p) \right) \tag{III.3}$$

– avec h_{CV}

$$ISRE_{CV}^{(j)} = \frac{1}{card(T_{grid})} \sum_{t_i \in T_{grid}} \log^2 \left(\widehat{q}_p^{(j)}(t_i, h_{CV}) / F_{t_i}^{-1}(p) \right) \tag{III.4}$$

– avec h_{adapt}

$$ISRE_{adapt}^{(j)} = \frac{1}{card(T_{grid})} \sum_{t_i \in T_{grid}} \log^2 \left(\widehat{q}_p^{(j)}(t_i, \widehat{h}_{adapt}) / F_{t_i}^{-1}(p) \right) \tag{III.5}$$

où $\widehat{q}_p^{(j)}(\cdot, h)$ désigne l'estimateur du quantile d'ordre p pour la $j$ème simulation. Pour comparer la qualité des 3 estimateurs, nous déterminons pour $N_{MC} = 1000$ simulations et pour chaque point $t \in T_{grid}$ les erreurs quadratiques moyennes relatives notées $MSRE$ (Mean Squared Relative Error) suivantes :

$$MSRE_{opt}(t,p) = \frac{1}{N_{MC}} \sum_{j=1}^{N_{MC}} \log^2 \left(\widehat{q}_p^{(j)}(t, h_{opt}) / F_t^{-1}(p) \right),$$

$$MSRE_{CV}(t,p) = \frac{1}{N_{MC}} \sum_{j=1}^{N_{MC}} \log^2 \left(\widehat{q}_p^{(j)}(t, h_{CV}) / F_t^{-1}(p) \right),$$

$$MSRE_{adapt}(t,p) = \frac{1}{N_{MC}} \sum_{j=1}^{N_{MC}} \log^2 \left(\widehat{q}_p^{(j)}(t, \widehat{h}_{adapt}) / F_t^{-1}(p) \right).$$

Tout d'abord, nous analysons les résultats calculés pour $n = 50000$ et $p = 0.999$. Pour un échantillon simulé, la Figure III.8 représente simultanément le logarithme de l'estimateur $\widehat{q}_p^{(1)}(\cdot, h_{opt})$, $\widehat{q}_p^{(1)}(\cdot, h_{CV})$ et $\widehat{q}_p^{(1)}(\cdot, \widehat{h}_{adapt})$ (en trait pointillé) et le 0.999-quantile théorique (en trait plein) en fonction de t. Comme attendu, le choix de $h = h_{opt}$ donne les meilleurs résul-

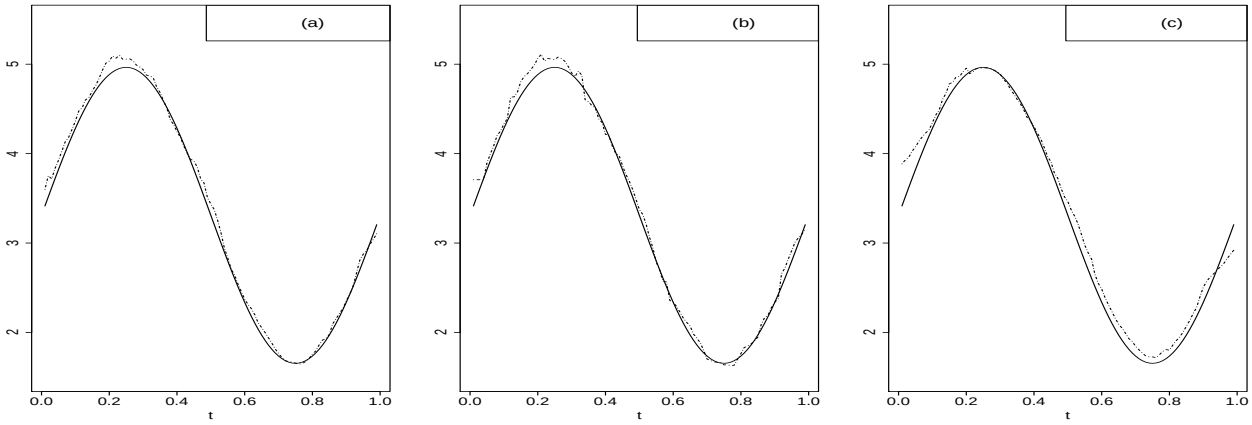

Figure III.8 – Représentation pour un échantillon simulé du logarithme des estimateurs de $F_t^{-1}(0.999)$ (en trait pointillé) et du 0.999-quantile théorique (en trait plein) en fonction de t en choisissant la taille de la fenêtre h par (a) le choix $h = h_{opt} = 0.051$ et $ISRE^{(1)}_{opt} = 0.0068$; (b) la méthode adaptative $ISRE^{(1)}_{adapt} = 0.0076$; et (c) la méthode de la validation croisée $h_{CV} = 0.092$ et $ISRE^{(1)}_{CV} = 0.0183$.

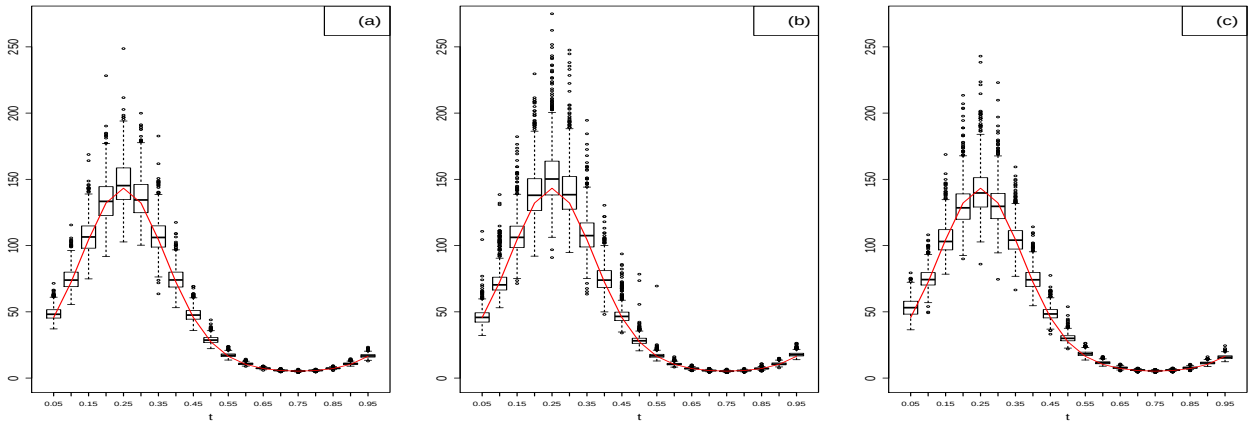

Figure III.9 – Boîte à moustaches des estimateurs de $F_t^{-1}(0.999)$ pour $N_{MC} = 1000$ en choisissant la taille de la fenêtre h par (a) le choix $h = h_{opt}$; (b) la méthode adaptative et (c) la méthode de la validation croisée.

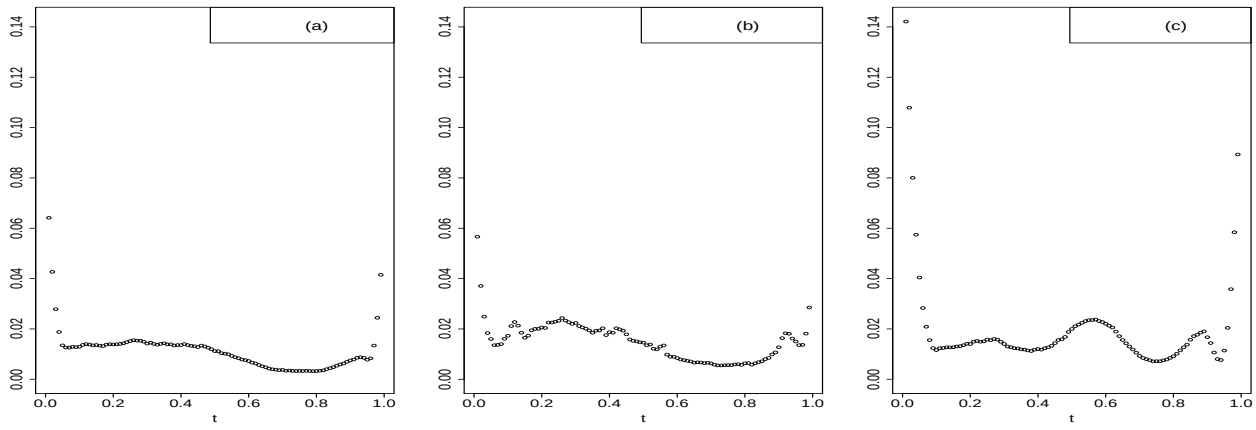

Figure III.10 – Représentation de l'erreur quadratique moyenne relative des estimateurs de $F_t^{-1}(0.999)$ en choisissant la taille de la fenêtre h par (a) le choix $h = h_{opt}$; (b) la méthode adaptative et (c) la méthode de la validation croisée.

tats au sens du critère $ISRE$ qui prend la plus petite valeur sur les trois choix de h. Pour $N_{MC} = 1000$ simulations, la Figure III.11 représente les erreurs relatives intégrées des estimateurs des quantiles $ISRE_{opt}^{(j)}$, $ISRE_{adapt}^{(j)}$ et $ISRE_{CV}^{(j)}$ pour $j = 1, \ldots, N_{MC}$. Nous remarquons que significativement en moyenne $ISRE_{h_{opt}} < ISRE_{adapt} < ISRE_{CV}$ (valeur de $p < 0.001$). De plus, la loi des erreurs obtenues par validation croisée a une queue plus lourde à droite que la loi des erreurs obtenue par la méthode adaptative et par le choix optimal $h = h_{opt}$. Au sens du critère de l'erreur relative intégrée, le choix $h = h_{opt}$ donne donc de meilleurs résultats que les deux autres méthodes. D'autre part, la méthode adaptative est sensiblement meilleure que la méthode de la validation croisée.

La Figure III.9 représente les boîtes à moustaches pour $N_{MC} = 1000$ échantillons simulés des estimateurs $\widehat{q}_{0.999}^{(j)}(\cdot, h_{opt})$, $\widehat{q}_{0.999}^{(j)}(\cdot, h_{CV})$ et $\widehat{q}_{0.999}^{(j)}(\cdot, \widehat{h}_{adapt})$ pour $t \in T_{grid}$. Nous observons un bon ajustement de ces estimateurs. La Figure III.10 représente les erreurs quadratique moyennes relatives $MSRE_{opt}(\cdot, 0.999)$, $MSRE_{adapt}(\cdot, 0.999)$ et $MSRE_{CV}(\cdot, 0.999)$ calculées aussi sur $N_{MC} = 1000$ simulations pour chaque point $t \in T_{grid}$. Nous pouvons observer des

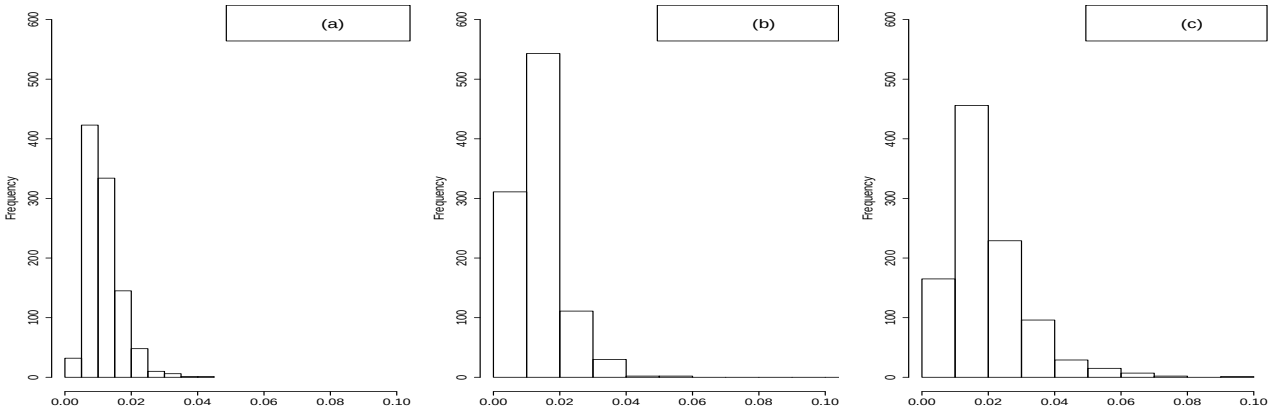

Figure III.11 – Représentation des histogrammes des erreurs relatives intégrées des estimateurs des quantiles $ISRE_{opt}^{(j)}$, $ISRE_{adapt}^{(j)}$ et $ISRE_{CV}^{(j)}$ calculées sur $N_{MC} = 1000$ simulations en choisissant la taille de la fenêtre h par (a) le choix $h = h_{opt}$; (b) la méthode adaptative et (c) la méthode de la validation croisée.

effets de bord pour les trois méthodes de choix de h. Les effets de bord sont moins marqués pour la méthode adaptative et pour le choix optimale $h = h_{opt}$. Pour les points t_i situés au centre de l'intervalle étudié, les erreurs quadratiques moyennes relatives sont faibles. Pour préciser, nous analysons maintenant les résultats pour plusieurs tailles d'échantillon n et plusieurs valeurs de p.

Les tableaux III.4, III.5 et III.6 représentent les erreurs quadratiques moyennes relatives calculés $MSRE_{opt}(\cdot, p)$, $MSRE_{adapt}(\cdot, p)$ et $MSRE_{CV}(\cdot, p)$ pour $t = 0.05, 0.10, \ldots, 0.95$, et $p = 0.99, 0.999, 0.9999$ avec les tailles d'échantillon $n = 5000$ et $n = 50000$. En comparant ces trois méthodes de choix de h, nous remarquons que le choix optimal $h = h_{opt}$ donne les meilleurs résultats mais en pratique il n'est pas possible d'utiliser cette méthode car la fonction F_t est inconnue. Pour $p = 0.99$, la méthode de la validation croisée est sensiblement meilleure que la méthode adaptative. Pour $p = 0.999$ et $p = 0.9999$, la méthode adaptative est sensiblement meilleure que la méthode de la validation croisée. Pour chaque valeur d'ordre p, il apparait que

Tableau III.4 – Erreur quadratique moyenne relative ($N_{MC} = 1000$) de l'estimateur $\widehat{q}_{0.99}(t)$ pour $t = 0.05, 0.10, \ldots, 0.95$ en considérant le modèle de Hall avec $n = 5000$ et $n = 50000$

	$n = 5000$			$n = 50000$		
t	Optimale	CP	CV	Oracle	CP	CV
0.05	0.0512	0.1297	0.0711	0.0040	0.0055	0.0072
0.10	0.0263	0.0858	0.0338	0.0042	0.0059	0.0038
0.15	0.0242	0.0690	0.0329	0.0045	0.0052	0.0042
0.20	0.0246	0.0728	0.0361	0.0047	0.0064	0.0045
0.25	0.0239	0.0712	0.0321	0.0051	0.0072	0.0053
0.30	0.0227	0.0700	0.0316	0.0048	0.0069	0.0047
0.35	0.0211	0.0689	0.0295	0.0049	0.0065	0.0043
0.40	0.0194	0.0619	0.0257	0.0043	0.0070	0.0040
0.45	0.0199	0.0562	0.0266	0.0040	0.0040	0.0039
0.50	0.0214	0.0601	0.0268	0.0034	0.0038	0.0039
0.55	0.0196	0.0538	0.0277	0.0028	0.0036	0.0038
0.60	0.0186	0.0361	0.0250	0.0023	0.0026	0.0033
0.65	0.0121	0.0220	0.0176	0.0015	0.0021	0.0025
0.70	0.0063	0.0143	0.0103	0.0013	0.0021	0.0017
0.75	0.0046	0.0114	0.0072	0.0012	0.0018	0.0015
0.80	0.0072	0.0132	0.0104	0.0012	0.0021	0.0016
0.85	0.0140	0.0228	0.0174	0.0016	0.0023	0.0025
0.90	0.0188	0.0304	0.0186	0.0023	0.0037	0.0032
0.95	0.0573	0.0432	0.0579	0.0024	0.0044	0.0028

Tableau III.5 – Erreur quadratique moyenne relative ($N_{MC} = 1000$) de l'estimateur $\widehat{q}_{0.999}(t)$ pour $t = 0.05, 0.10, \ldots, 0.95$ en considérant le modèle de Hall avec $n = 5000$ et $n = 50000$

	$n = 5000$			$n = 50000$		
t	Optimale	CP	CV	Optimale	CP	CV
0.05	0.1502	0.3622	0.3485	0.0135	0.0160	0.0404
0.10	0.0796	0.2484	0.4480	0.0129	0.0173	0.0116
0.15	0.0677	0.2058	0.5000	0.0136	0.0165	0.0127
0.20	0.0688	0.2143	0.5816	0.0138	0.0206	0.0140
0.25	0.0653	0.2202	0.6087	0.0152	0.0233	0.0158
0.30	0.0626	0.2104	0.5537	0.0142	0.0224	0.0140
0.35	0.0594	0.2015	0.4985	0.0143	0.0185	0.0119
0.40	0.0561	0.1751	0.4175	0.0136	0.0187	0.0121
0.45	0.0587	0.1803	0.3343	0.0128	0.0179	0.0144
0.50	0.0670	0.1914	0.2432	0.0112	0.0146	0.0200
0.55	0.0597	0.1651	0.1635	0.0093	0.0130	0.0235
0.60	0.0587	0.1096	0.1077	0.0073	0.0084	0.0221
0.65	0.0396	0.0681	0.0676	0.0048	0.0066	0.0156
0.70	0.0204	0.0451	0.0453	0.0037	0.0063	0.0093
0.75	0.0149	0.0362	0.0396	0.0033	0.0056	0.0072
0.80	0.0230	0.0414	0.0445	0.0033	0.0064	0.0091
0.85	0.0434	0.0697	0.0663	0.0048	0.0071	0.0157
0.90	0.0580	0.0937	0.1049	0.0075	0.0127	0.0167
0.95	0.1636	0.1382	0.1813	0.0078	0.0150	0.0114

Tableau III.6 – Erreur quadratique moyenne relative ($N_{MC} = 1000$) de l'estimateur $\widehat{q}_{0.9999}(t)$ pour $t = 0.05, 0.10, \ldots, 0.95$ en considérant le modèle de Hall avec $n = 5000$ et $n = 50000$

	$n = 5000$			$n = 50000$		
t	Optimale	CP	CV	Optimale	CP	CV
0.05	0.3011	0.7131	0.9795	0.0293	0.0323	0.1227
0.10	0.1615	0.4972	1.2737	0.0274	0.0351	0.1082
0.15	0.1344	0.4184	1.3988	0.0284	0.0357	0.1032
0.20	0.1331	0.4336	1.7252	0.0278	0.0438	0.1177
0.25	0.1295	0.4553	1.8152	0.0308	0.0492	0.1230
0.30	0.1235	0.4291	1.6557	0.0293	0.0475	0.1130
0.35	0.1171	0.4098	1.4906	0.0298	0.0384	0.1075
0.40	0.1148	0.3562	1.2107	0.0280	0.0465	0.0868
0.45	0.1197	0.3812	0.8860	0.0270	0.0430	0.1047
0.50	0.1387	0.3986	0.6513	0.0238	0.0330	0.0909
0.55	0.1229	0.3378	0.4099	0.0197	0.0283	0.0882
0.60	0.1202	0.2235	0.2755	0.0152	0.0178	0.0703
0.65	0.0839	0.1401	0.1673	0.0101	0.0138	0.0504
0.70	0.0439	0.0941	0.1235	0.0077	0.0130	0.0362
0.75	0.0318	0.0755	0.1103	0.0068	0.0118	0.0307
0.80	0.0473	0.0858	0.1325	0.0066	0.0132	0.0353
0.85	0.0893	0.1428	0.1736	0.0101	0.0148	0.0483
0.90	0.1192	0.1925	0.2748	0.0160	0.0274	0.0577
0.95	0.3263	0.2884	0.4274	0.0163	0.0321	0.0713

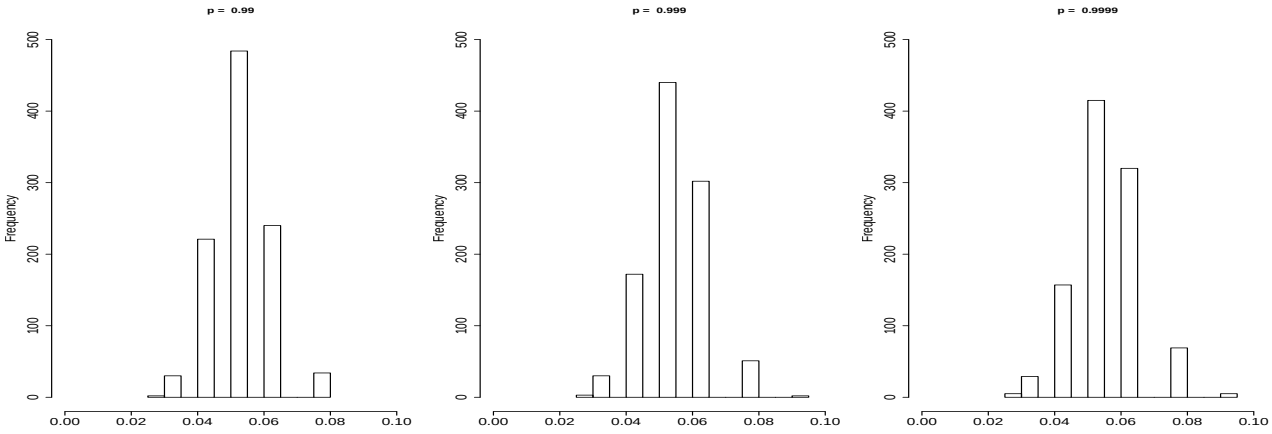

Figure III.12 – Histogramme des tailles de fenêtres estimées pour le choix optimal $h = h_{opt}$ sur $N_{MC} = 1000$ échantillons de modèle de mélange de taille $n = 50000$ pour $p = 0.99, 0.999, 0.9999$.

plus la queue de F_t est légère plus MSRE est faible. Pour chaque valeur de t, plus l'ordre p est élevé plus MSRE est élevée. Par conséquent, il est plus difficile d'estimer $F_t^{-1}(p)$ quand p est élevé. De ces résultats, nous remarquons que l'estimateur de la taille de la fenêtre par la méthode adaptative n'est pas affecté par l'ordre p du quantile choisi tandis que par la méthode de la validation croisée nous obtenons de moins bons résultats pour une valeur de p élevé.

Les Figures III.12, III.13 et III.14 représentent les histogrammes des tailles de fenêtres estimées par le choix optimal $h = h_{opt}$, la méthode de la validation croisée et la méthode adaptative (aux points $t = 0.1, \ldots, 0.9$) sur $N_{MC} = 1000$ échantillons de taille échantillon $n = 50000$. Nous remarquons que les tailles de fenêtres choisies par le choix optimal $h = h_{opt}$ restent stable par rapport au changement de l'ordre p.

Dans ces simulations, les temps de calcul obtenus par la méthode adaptative sont les plus longs et les temps calculs avec le choix optimal h_{opt} et la validation croisée sont presque identiques. Pour un échantillon de taille $n = 50000$, il faut approximativement 16 minutes par la méthode adaptative et 5 minutes pour la méthode optimale et la validation croisée (Processeur

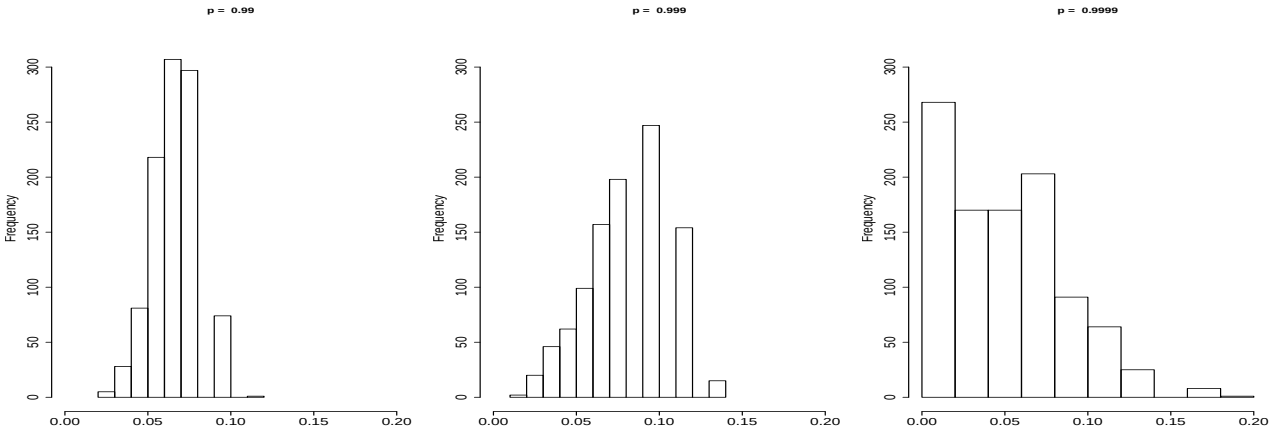

Figure III.13 – Histogramme des tailles de fenêtres estimées par la méthode de la validation croisée sur $N_{MC} = 1000$ échantillons de modèle de mélange de taille $n = 50000$ pour $p = 0.99, 0.999, 0.9999$.

Intel(R) Core(TM) $i5 - 2400$ CPU@ 3.10 GHz 3,1 GHz).

4 Modèle de Burr avec données dépendantes

Nous nous intéressons ici au comportement de notre procédure pour des données dépendantes. Afin de simuler des données dépendantes, nous suivons l'article de Gardes et Girard [58]. Considérons le modèle de Burr donné par :

$$F_t(x) = 1 - \left(1 + x^{-\rho/\gamma_t}\right)^{1/\rho}, \qquad x > 0, \rho > 0, 0 \leq t \leq 1, \tag{III.6}$$

où

$$\gamma_t = \frac{1}{2}\left(\frac{1}{10} + \sin(\pi t)\right)\left(\frac{11}{10} - \frac{1}{2}\exp(-64(t-1/2)^2)\right), \qquad 0 \leq t \leq 1 \tag{III.7}$$

En utilisant le modèle (III.6), on génère $N = 1000$ réplications d'un échantillon $\{X_{t_i} : i = 1, ..., n\}$ de taille $n = 5000$ où $\{t_i = i/n : i = 1, ..., n\}$ est un design fixe sur l'intervalle $[0, 1]$. La procédure de simulation des échantillons décrite dans (Fawcett et Walshaw [59], Gardes et Girard [58]) est la suivante :

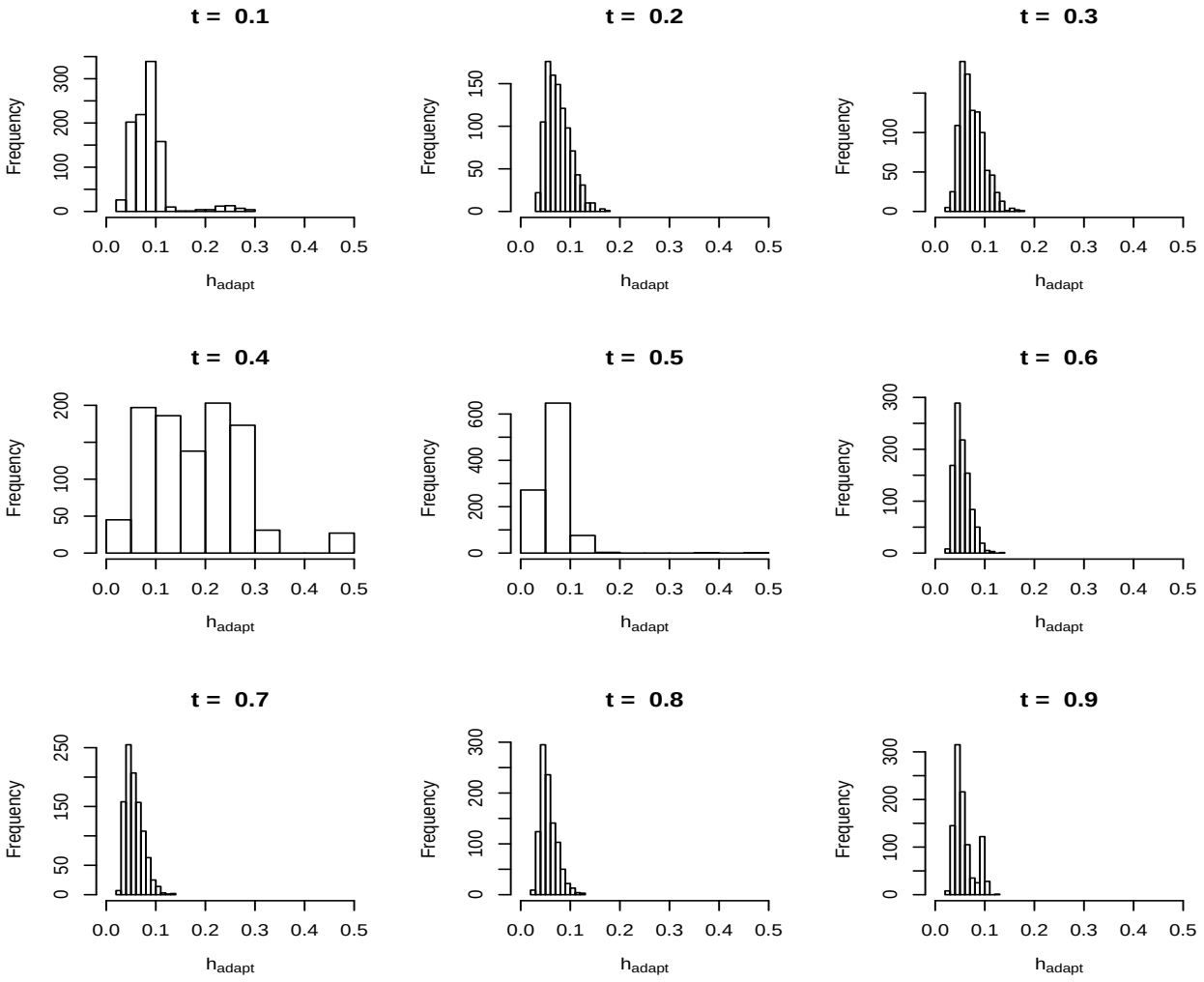

Figure III.14 – Histogramme des tailles de fenêtres estimées par la méthode adaptative sur $N_{MC} = 1000$ échantillons de modèle de mélange de taille $n = 50000$ aux points $t = 0.1, \ldots, 0.9$.

1. Générer Y_{t_1} suivant la loi de Fréchet standard.
2. Pour $i = 1, ..., n-1$, calcul de la loi conditionnelle de $Y_{t_{i+1}}$ sachant Y_{t_i} en utilisant le fait que la loi marginale de $(Y_{t_{i+1}}, Y_{t_i})$ est la loi bivariée des valeurs extrêmes donnée par

$$G(u,v) = \exp\left(V(u,v)\right),$$

où $V(u,v)$ est une fonction de dépendance de type logistique définie par :

$$V(u,v) = \left(u^{-1/\alpha} + v^{-1/\alpha}\right)^{\alpha}, u > 0, v > 0$$

où $\alpha \in (0,1]$. Le paramètre α contrôle ici la dépendance entre deux observations consécutives. Plus α est proche de 0, plus les deux observations consécutives sont dépendantes. Le cas $\alpha = 1$ correspond au cas des données indépendantes et $\alpha \to 0$ au cas des données dépendantes complétées.

3. Calcul de $X_{t_i} = F_{t_i}^{-1}\left(F_{Y_{t_i}}(Y_{t_i})\right)$ pour $i = 1, \ldots, n$ où $F_t^{-1}(.)$ désigne la fonction inverse généralisée de $F_t(.)$ et $F_{Y_{t_i}}(.)$ la fonction de répartition de Y_{t_i} de loi de Fréchet standard

$$F_{Y_{t_i}}(y) = 1_{\{y>0\}}(y) \exp\left(-\frac{1}{y}\right).$$

Dans un premier temps, nous présentons les résultats de nos simulations pour $\alpha = 0.8$ et $n = 50000$. Dans un second temps, nous donnerons les résultats pour différentes valeurs de $\alpha \in \{0.4, 0.6, 0.8, 1\}$ et $n = 5000$.

Les tableaux III.7 et III.8 représentent l'erreur quadratique moyenne relative pour plusieurs valeur de $\alpha \in \{0.4, 0.6, 0.8, 1.0\}$. Ces résultats montrent que :

- les erreurs quadratiques moyennes relatives restent petites même dans le cas de dépendance (voir moyennes dans les Tableaux III.7 et III.8),
- la méthode optimale donne comme attendu les meilleurs résultats mais cette méthode ne peut s'appliquer en pratique car la fonction quantile $F_t^{-1}(p)$ est inconnue,

Tableau III.7 – Erreur quadratique moyenne ($N_{MC} = 1000$) de l'estimateur $\widehat{q}_{0.99}(t)$ pour $t = 0.05, 0.10, \ldots, 0.95$ en considérant le modèle de Burr avec $\alpha = 0.4$, $\alpha = 0.6$ et $n = 5000$

	$\alpha = 0.4$			$\alpha = 0.6$		
t	Optimale	CP	CV	Optimale	CP	CV
0.05	0.0727	0.0343	0.0304	0.0471	0.0182	0.0217
0.10	0.0723	0.0728	0.0680	0.0341	0.0445	0.0403
0.15	0.0896	0.1224	0.1119	0.0435	0.0678	0.0599
0.20	0.1181	0.1926	0.1708	0.0591	0.0983	0.0723
0.25	0.1545	0.2589	0.2590	0.0703	0.1420	0.1188
0.30	0.1650	0.2868	0.2919	0.0780	0.1596	0.1735
0.35	0.1682	0.2688	0.2666	0.0813	0.1765	0.1722
0.40	0.1392	0.2418	0.2294	0.0699	0.1317	0.1202
0.45	0.1032	0.1712	0.1612	0.0625	0.1022	0.0841
0.50	0.1086	0.1518	0.1453	0.0613	0.0724	0.0591
0.55	0.1110	0.1974	0.1719	0.0533	0.0868	0.0814
0.60	0.1309	0.2619	0.2463	0.0606	0.1223	0.1165
0.65	0.1612	0.2644	0.2643	0.0817	0.1599	0.1382
0.70	0.1514	0.2643	0.2630	0.0820	0.1450	0.1395
0.75	0.1159	0.2175	0.2299	0.0603	0.1257	0.1055
0.80	0.0977	0.1675	0.1614	0.0443	0.0924	0.0738
0.85	0.0833	0.1283	0.1231	0.0632	0.0775	0.0772
0.90	0.0677	0.0788	0.0703	0.0291	0.0346	0.0388
0.95	0.0889	0.0408	0.0414	0.0390	0.0158	0.0188
Moyenne	0.1158	0.1801	0.174	0.059	0.0986	0.0901
Erreur type	0.0076	0.0188	0.0191	0.0037	0.0112	0.0106

– Les erreurs quadratiques moyennes relatives sont d'autant plus petites que la valeur de α augmente donc que l'on se rapproche de l'indépendance.

Nous pouvons cependant remarquer contrairement aux résultats du paragraphe 3 que la méthode de la validation croisée donne ici des résultats sensiblement meilleurs que la méthode adaptative et ceci même dans le cas de l'indépendance ($\alpha = 1$). Cette légère différence en moyenne peut s'expliquer par le fait que la queue de distribution dans le cas de cette simulation est plus légère que dans la simulation du paragraphe 3.

Au bilan, les résultats obtenus sont encourageants et démontrent que notre procédure donne

Tableau III.8 – Erreur quadratique moyenne ($N_{MC} = 1000$) de l'estimateur $\widehat{q}_{0.99}(t)$ pour $t = 0.05, 0.10, \ldots, 0.95$ en considérant le modèle de Burr avec $\alpha = 0.8$, $\alpha = 1.0$ et $n = 5000$

	$\alpha = 0.8$			$\alpha = 1.0$		
t	Optimale	CP	CV	Optimale	CP	CV
0.05	0.0218	0.0074	0.0154	0.0118	0.0044	0.0165
0.10	0.0147	0.0179	0.0168	0.0085	0.0096	0.0102
0.15	0.0191	0.0321	0.0240	0.0110	0.0179	0.0126
0.20	0.0262	0.0490	0.0346	0.0153	0.0237	0.0172
0.25	0.0311	0.0654	0.0465	0.0188	0.0303	0.0229
0.30	0.0364	0.0732	0.0642	0.0218	0.0326	0.0276
0.35	0.0410	0.0795	0.0597	0.0215	0.0395	0.0284
0.40	0.0356	0.0690	0.0578	0.0192	0.0304	0.0232
0.45	0.0330	0.0467	0.0368	0.0175	0.0370	0.0203
0.50	0.0314	0.0356	0.0348	0.0167	0.0222	0.0218
0.55	0.0315	0.0485	0.0394	0.0188	0.0364	0.0212
0.60	0.0355	0.0718	0.0508	0.0205	0.0321	0.0230
0.65	0.0474	0.0874	0.0644	0.0275	0.0450	0.0342
0.70	0.0392	0.0706	0.0586	0.0204	0.0319	0.0255
0.75	0.0319	0.0643	0.0452	0.0182	0.0285	0.0221
0.80	0.0244	0.0493	0.0352	0.0167	0.0258	0.0181
0.85	0.0199	0.0419	0.0258	0.0110	0.0172	0.0132
0.90	0.0171	0.0186	0.0173	0.0088	0.0101	0.0102
0.95	0.0228	0.0076	0.0168	0.0121	0.0042	0.0168
Moyenne	0.0295	0.0493	0.0392	0.0166	0.0252	0.0203
Erreur type	0.002	0.0056	0.0039	0.0012	0.0027	0.0015

des résultats satisfaisants (faibles valeurs des erreurs quadratiques moyennes relatives) en pratique même dans le cas de la dépendance.

5 Conclusion

En conclusion au vu des résultats, ces simulations nous ont permis non seulement d'illustrer la pertinence des résultats théoriques mais aussi de montrer que la procédure développée possède les trois avantages suivants : 1) elle reste stable par rapport au changement de modèle. Cette remarque est à nuancer dans le sens ou nous avons testé seulement trois modèles ; 2) elle donne des résultats satisfaisants pour les deux méthodes de choix de la taille de la fenêtre (validation croisée et adaptative) en comparaison avec une méthode optimale ; 3) elle reste relativement peu sensible à la dépendance induite par le paramètre $\alpha \in [0, 1]$.

Chapitre IV

Application

Ce paragraphe concerne un projet d'article à soumettre avec des collègues biologistes de l'université de Bretagne Sud et de l'université de Bordeaux.

Les activités humaines sont responsables d'importants rejets d'agents polluants dans le milieu naturel. Ces polluants entraînent la dégradation de nombreux biotopes, perturbant les écosystèmes et posant également des problèmes en termes de santé publique. Des réglementations et des contrôles sur la qualité des eaux ont été mis en place. Parmi ces contrôles, les bioindicateurs sont de plus en plus utilisés car ils peuvent se révéler très efficaces par leurs capacités à révéler la présence de traces (concentrations très faibles) de contaminant. Nous utilisons ici comme moyen de surveillance du milieu la valvométrie. La valvométrie (mesure de l'activité des valves de mollusques) est une technique qui permet d'enregistrer les réactions de bivalves, face aux changements de la qualité de l'eau dans laquelle ils vivent. Le premier enregistrement publié de l'activité valvaire de mollusques bivalves date de 1909 (Marceau, 1909 [1]). Depuis, des dizaines de systèmes ont été développés. Les deux plus connus, parce qu'ils ont trouvé une application commerciale, sont le Mosselmonitor (Kramer et Foekema, 2001 [2]) et le Dreissena monitor (Borcherding, 2006 [60]). L'Institut Français de Recherche pour l'Exploitation de la MER (IFREMER) a aussi développé un système avec la société Micrel NKE. Un système a aussi été développé au sein de l'équipe scientifique de l'oeil du mollusque composée de Pierre Ciret, Gilles Durrieu, Jean-Charles Massabuau, Mohamedou Sow et Damien Tran pour plusieurs raisons non satisfaites par ces systèmes. L'objectif était de ne pas travailler avec

des animaux collés mais sur des animaux libres de tous mouvements. Des animaux stressés par des contraintes expérimentales peuvent en effet présenter une perte de sensibilité très importante. Notre objectif était aussi de caractériser et traiter statistiquement le signal obtenu et le traiter sur de longues séries. Un des buts est de faire du suivi à très long terme afin (1) de comprendre le fonctionnement et la trajectoire temporelle des systèmes environnementaux, (2) d'appréhender leur résilience, leurs réponses à des évènements perturbateurs et leur dynamique d'ajustement.

Le dispositif développé est une technologie robuste supportant une intervention par an ou moins, et disposant d'un accès web avec mise à jour journalière des résultats. Le principe du biocapteur valvométrique développé est le suivant. Nous collons sur les valves des bivalves des électrodes (électro-aimant) légères (200 mg hors enrobage) et non invasives (l'animal est libre de se positionner et se déplacer) qui permettent de mesurer en permanence leur état d'ouverture. Ces électrodes sont gérées par une carte analogique immergée à coté des animaux en boitier étanche. Le tout est relié à une seconde carte électronique en surface pilotée sous Linux. Les données brutes sont ensuite transférées tous les jours par le réseau de la téléphonie mobile (GPRS) et internet (FTP) sur des stations de travail. La fréquence d'échantillonnage est fixée à une mesure toutes les 0.1s mais cette fréquence peut être changé. Le tout consomme seulement 1 watt et est alimentable par des panneaux solaires, des batteries ou une source de courant classique. Pour avoir un nombre d'animaux représentatif, nous travaillons sur des groupes de 16 animaux. Nous disposons donc d'un total de 864000 couples de points par jour tous les jours. Une fois les données brutes arrivées sur la station de travail, elles sont modélisées et traitées statistiquement. Deux heures après le transfert depuis le terrain, les résultats sont ensuite mis en accès public et/ou professionnel sur le site web (l'Oeil du Mollusque http://molluscan-eye.epoc.u-bordeaux1.fr/). Plusieurs sites sont actuellement équipés de ce système à différents endroits dans le monde (Espagne, France, Norvège, Nouvelle Calédonie et

Russie). Le premier site français équipé est face à la Station de Biologie Marine d'Arcachon. Il fonctionne depuis mars 2006. Ce sont des huîtres qui sont utilisés pour détecter l'arrivée de contaminant dans le bassin d'Arcachon. Pour le 2ème site, nous avons installé 16 bénitiers dans le lagon sud de Nouvelle Calédonie. L'idée était de participer à la création d'un Observatoire de l'Environnement, en relation avec la province sud de Nouvelle Calédonie pour suivre l'impact potentiel d'une nouvelle mine de nickel et de cobalt. Nous considérons dans cette étude le troisième site équipé, celui de Locmariaquer en Bretagne Sud dans le golf du Morbihan installé depuis le 3 mars 2011 dans le cadre du projet ASPEET financé par l'Université de Bretagne Sud.

Le cas des mollusques bivalves est particulièrement intéressant en tant qu'espèce bioindicatrice car ce sont des animaux sédentaires qui sont spécifiquement témoins de changement locaux de la qualité de l'eau. Le but est de caractériser son activité valvaire (qui résume à elle seule l'ensemble du comportement locomoteur de ce type d'animal) en fonction des paramètres du milieu afin de discriminer ensuite un comportement atypique lié à une contamination aquatique. Au final, nous recherchons à caractériser une signature comportementale spécifique d'un milieu, de la nature d'un contaminant ou d'un facteur du milieu. Ces facteurs peuvent être soit des algues toxiques, soit des déchets de l'activité anthropique comme les métaux lourds, les pesticides ou les HAPs, soit lié au réchauffement global dans un fjord comme par exemple en Arctique à Ny- Alesund. Etant basé originalement à Arcachon, des études ont concerné en particulier les efflorescences d'algues toxiques produisant des toxines diarrhéiques (Diarrheic Shellfish Poisoning toxins) ou paralysantes (Paralytic Shell fish Poisoning toxins). Elles sont en très forte augmentation le long des côtes françaises et elles causent de lourdes pertes économiques dans le secteur conchylicole et ostréicole dues aux fermetures administratives de commercialisation des coquillages. Notre hypothèse de travail est que les organismes aquatiques doivent présenter des perturbations de leur comportement de base et de leurs rythmes biologiques face

aux contraintes du milieu. Les caractéristiques typiques de leurs activités physiologiques ou comportementales seront modifiées et seront le témoin de cette perturbation du milieu. Le rythme a une origine endogène (gènes circadiens) lié à la notion d'horloge biologique. Mais les synchroniseurs de ces rythmes, appelés zeitgeiber, sont d'origines exogènes comme la lumière ou la marée. Des premières études ont consisté dans un premier temps à comprendre d'abord comment les facteurs du milieu règlent la rythmicité biologique pour discriminer ensuite un comportement qui n'est pas régit par ces facteurs, mais résulte d'une perturbation du milieu ambiant. Notre approche est basée sur l'étude à haute fréquence du comportement valvaire de bivalves. Il nous faut aujourd'hui des moyens pour approfondir l'étude des réactions valvaires en fonction des variations des caractéristiques du milieu (physico-chimie, biologie). A ce sujet il faut savoir que nous possédons déjà des séries longues. D'une façon générale on considère que ces animaux sont de bons bioindicateurs de contaminations locales et de bon bioconcentrateurs de contaminants à cause de leur activité de filtration régulière de l'eau. Nos premières études expérimentales montrent une très bonne réactivité du comportement valvaire face à des contaminants métalliques comme le cadmium, le mercure, le cuivre et l'uranium (Tran et al. 2003, 2004, 2007 [4, 5, 49] ; Fournier et al., 2005 [61]), de parasites (Chambon et al., 2007 [6]) ou d'algues toxiques. Les réponses valvaires en milieu contrôlé, sont spécifiques à la nature du contaminant. Dans ce contexte, des approches de modélisation dose réponse ont permis d'estimer la concentration de contaminant qui a un effet sur les animaux (Tran et al., 2004 [5]).

Au niveau traitement mathématiques et statistique des données, nous pouvons citer les travaux de Azaïs, Coudret et Durrieu (2014) [11], Coudret, Durrieu et Saracco (2014) [10], Schmitt et al. (2011) [9], Sow et al. (2011) [8], Tran et al. (2003, 2004, 2007 et 2011) [4, 5, 49, 62], Schwartzmann et al. (2011) [7], Chambon et al. (2007) [6] , Liao et al. (2005) [63] et Liao et al. (2006) [64]. Les développements statistiques précédents touchent aux processus stochastiques (processus de renouvellement et Shot Noise), au quantile de régression et à la régression non

paramétrique basée sur des estimateurs récursifs ou non de la fonction de lien.

1 Données site Locmariaquer

Figure IV.1 – Électrodes sur un groupe d'huîtres (http ://molluscan-eye.epoc.u-bordeaux1.fr/). Crédit photo Jean-Charles Massabuau.

Les Figures IV.1 et IV.2 présentent respectivement un groupe d'huîtres équipé de ses électro-aimants et le système installé du biocapteur, à savoir l'huître, dont l'information est accessible sur internet. Dans le cadre du projet ASPPET, un valvomètre a été installé à Locmariaquer ($N47°34,075' \ / \ W2°55,970'$) situé à l'entrée du Golfe du Morbihan en Bretagne Sud. L'acquisition, le transfert et le traitement des données fonctionnent de manière automatique depuis le 3 mars 2011. Ce dispositif de suivi est installé au sein de concessions ostréicoles accessibles à pieds à marée basse. Les données brutes correspondant à l'activité d'ouverture et de fermeture des huîtres sont alors transférées tous les jours du terrain à une station de travail sous linux par le réseau de la téléphonie mobile (GPRS) et internet (FTP). Pour avoir un nombre d'animaux représentatif, nous travaillons sur un groupes de 16 animaux. Nous disposons donc d'un total de 864000 triplets (le numéro de l'huître, le temps et l'amplitude de l'ouverture) de points par

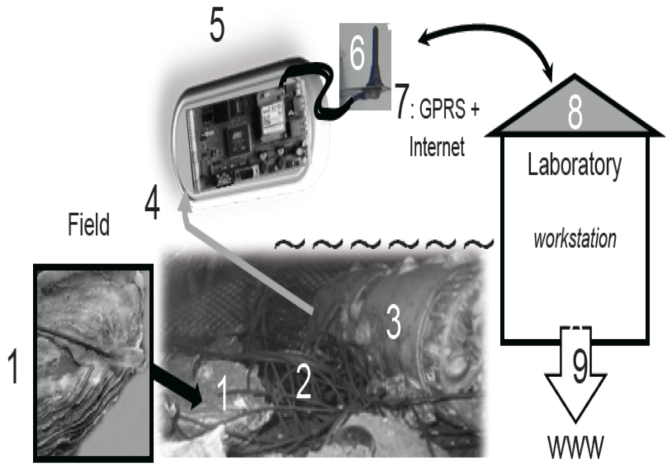

Figure IV.2 – Représentation synoptique du système installé du terrain au laboratoire : (1) une huître équipée de 2 électrodes ; (2) cables connectant les huîtres à la première carte électronique ; (3) la première carte électronique se trouve dans ce boitier étanche cylindrique ; (4) connexion électrique entre les 2 cartes électroniques ; (5) seconde carte électronique sous Linux avec carte Sim ; (6) Antenne GPRS ; (7) Connections GPRS et internet ; (8) Laboratoire ; (9) Mise à jour tous les jours sur internet.

jour. Quelques résultats sont mis automatiquement en accès public et professionnel sur un site web (l'Oeil du Mollusque http://molluscan-eye.epoc.u-bordeaux1.fr).

La Figure IV.3 correspond à l'activité d'une huître sur une journée à Locmariaquer. On peut remarquer dans ce signal 3 états caractéristiques du comportement des animaux qui sont des activités de fermeture, d'ouverture et fermetures et ouvertures partielles.

2 Résultats

A partir des données collectées, nous déterminons numériquement les vitesses d'ouvertures et fermetures des animaux. La vitesse de fermeture et ouverture des huîtres fournit une indication importante du changement de comportement de l'animal comme par exemple une ponte ou

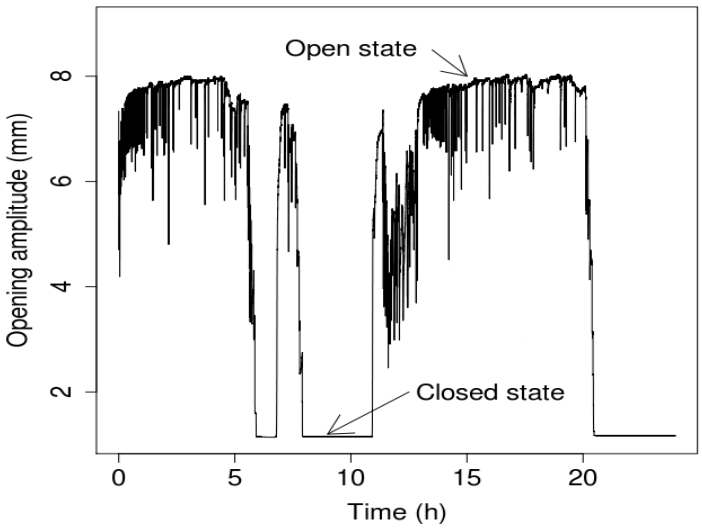

Figure IV.3 – Données valvométrique associées à l'activité d'une huître à Locmariaquer.

un état de fébrilité caractérisé par des fermetures-ouvertures partielles rapides anormales liées à un stress ou à une perturbation environnementale. Dans un milieu perturbé ou en situation de ponte par exemple, il est commun d'observer des vitesses de fermetures et ouvertures partielles rapides. Du point de vue des évènements extrêmes, nous nous intéressons à la détection de comportements extrêmes liés à des perturbations. Pour cela, nous appliquons l'approche statistique développée dans la partie I dans le but d'extraire des perturbations extrêmes et d'aider à la surveillance d'un écosystème marin. Les mesures sont collectées par les deux cartes électroniques de la Figure IV.2 qui gèrent à la fois l'acquisition (toutes les 1.6 secondes) sur les 16 huîtres et le transfert des données.

Nous appliquons pour un quantile $p = 0.999$ ici la méthode statistique présentée dans la partie I sur les données du 4 mars 2011 au 21 août 2011 à Locmariaquer. La figure IV.4 montre sur la journée du 18 avril 2011 un très bon comportement de l'estimateur de la probabilité de dépasser un seuil fixé *a priori* ainsi que du 0.999-quantile. Pour faciliter la visualisation des 0.999-quantiles extrêmes des vitesses sur les 16 huîtres pour la période du 4 mars (63ème jour

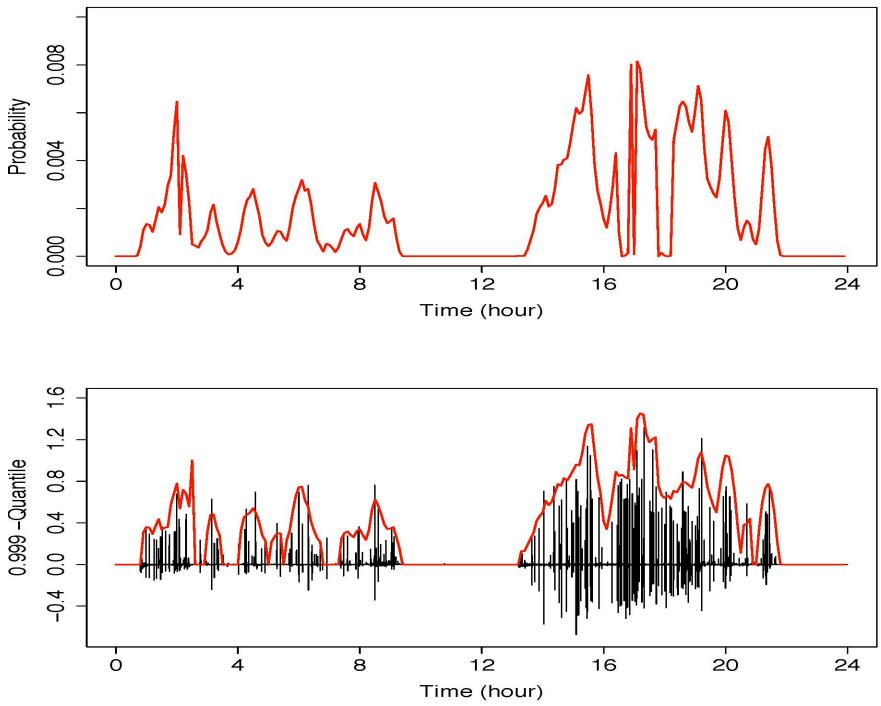

Figure IV.4 – La ligne rouge sur la Figure du haut représente l'estimateur de probabilité de $P(X_t > 0.3)$ le 18 avril 2011. La ligne rouge sur la figure du bas représente l'estimateur du 0.999-quantile le même jour. Les lignes en noir représentent les vitesses de fermeture.

de l'année) au 21 août 2011 (125ème jour de l'année), nous représentons graphiquement dans les Figures IV.5 (h estimée par la méthode de la validation croisée) et IV.6 (h estimée par la méthode adaptative) les 0.999-quantiles estimés par 3 classes de couleurs (gris, jaune et rouge) pour chaque instant $t \in [0, 24]$. Les couleurs grise, jaune et rouge correspondent respectivement aux 3 classes associées aux valeurs les plus faibles, aux valeurs intermédiaires et aux valeurs les plus élevées des estimateurs des 0.999-quantiles.

Ainsi pour chaque jour en ordonnée et pour chaque $t \in [0, 24]$ en abscisse, nous indiquons 16 lignes de points colorés par l'une des trois couleurs associées aux 3 classes. L'avantage de cette représentation est de fournir une visualisation simple de l'ensemble des résultats sur un

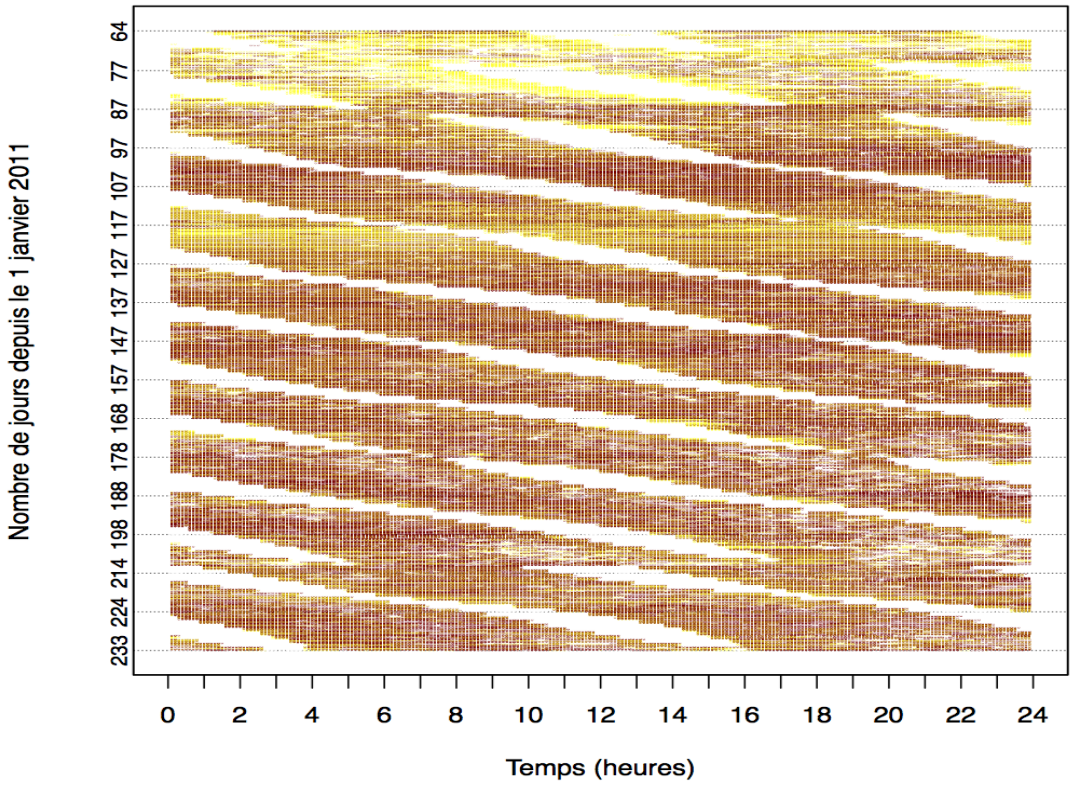

Figure IV.5 – Représentation du gradient (du gris au rouge) des 0.999-quantiles estimés pour l'ensemble des données acquises sur le site de Locmariaquer du 3 mars au 21 août. La taille de la fenêtre est estimée par la méthode de la validation croisée.

seul graphique. Nous observons des résultats similaires pour les 2 méthodes d'estimation de la fenêtre h et ces mois d'enregistrements nous ont permis de mettre en évidence des rythmes biologiques liés aux rythmes des marées chez l'huître (zones blanches sur la Figure IV.5 ou sur la Figure IV.6). Ce résultat est aussi confirmé en utilisant des modèles de régression non paramétrique dans [8, 10]. Cette représentation nous a aussi permis d'extraire visuellement des modifications importantes de l'activité des animaux. Nous remarquons en particulier une zone jaune (du 110 au 125ème jours) expliquée par un changement brutale des températures obser-

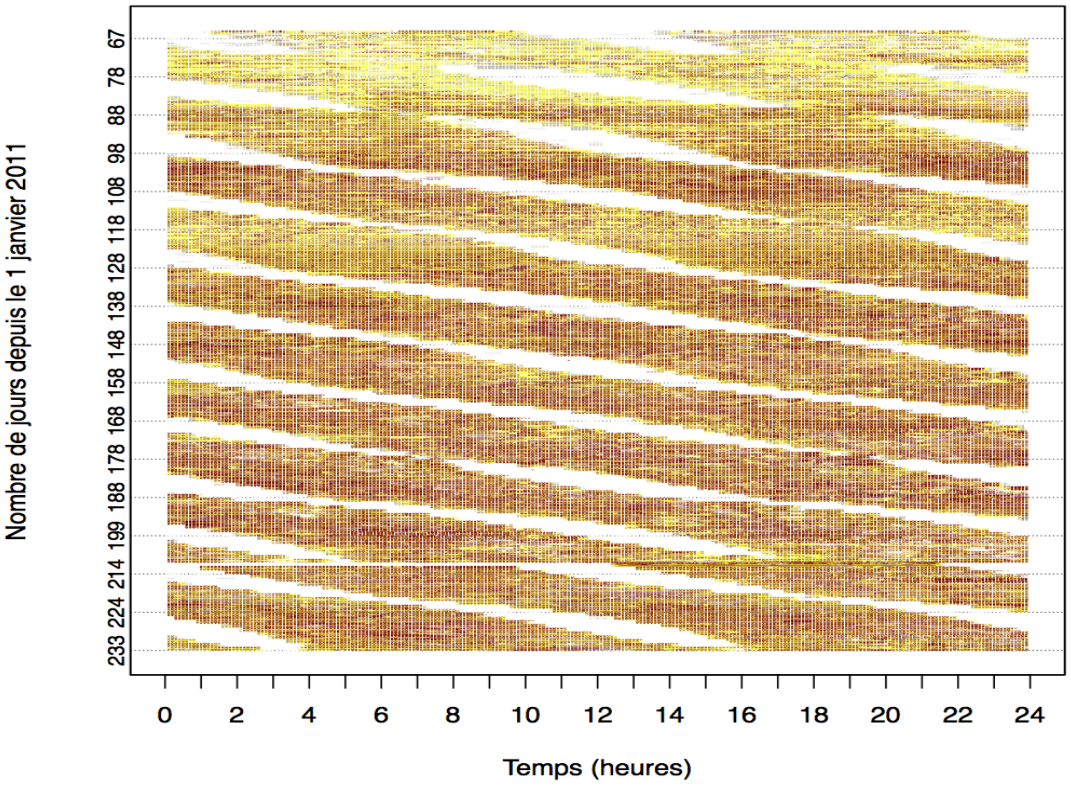

Figure IV.6 – Représentation du gradient (du gris au rouge) des 0.999-quantiles estimés pour l'ensemble des données acquises sur le site de Locmariaquer du 3 mars au 21 août. La taille de la fenêtre est estimée par la méthode adaptative.

vées de manière indépendante par un capteur de température installé juste à coté des huîtres et la zone la plus rouge correspondant à une activité plus intense des animaux liés à des activités notamment de reproduction (aux environs du 214ème jour). Afin de confirmer l'effet du changement de la température moyenne en quelques jours sur le comportement des animaux, nous représentons l'évolution des températures moyennes dans la Figure IV.7 et l'évolution moyenne des 0.999-quantiles estimés dans la Figure IV.8 sur la même période de temps. Nous remarquons l'existence de 4 phases qui sont associées aux modifications de paramètres environnementaux.

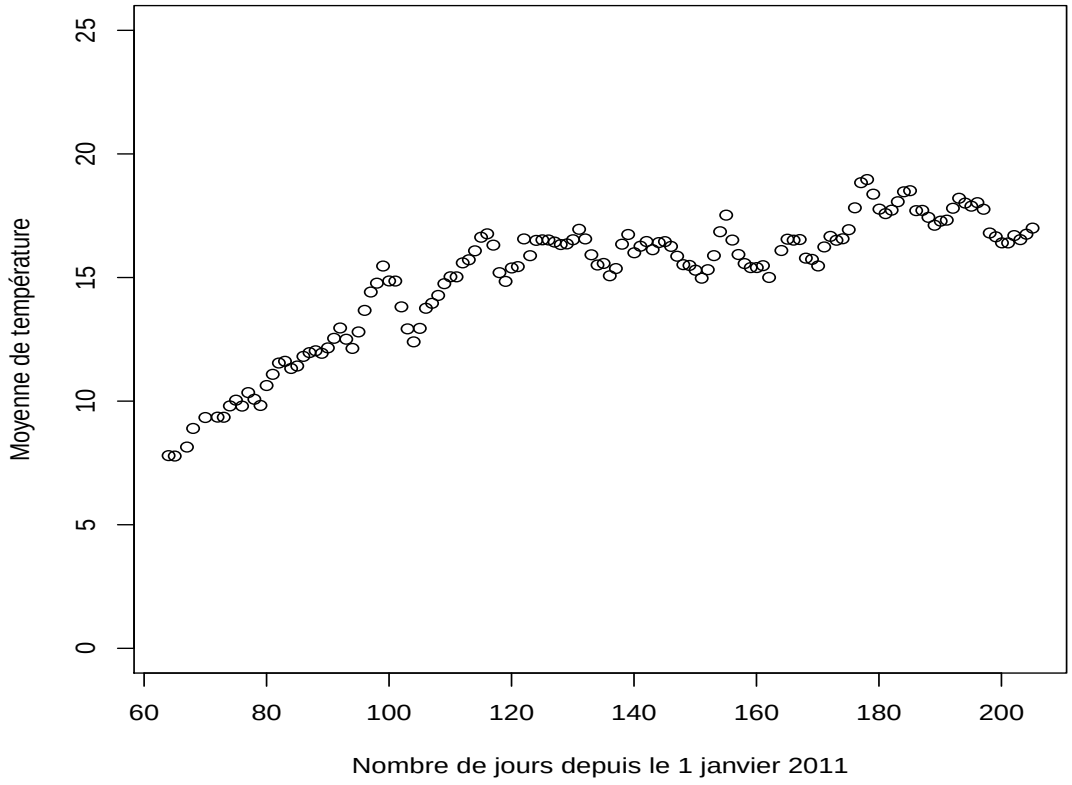

Figure IV.7 – Représentation de l'évolution de la température entre le 3 mars et le 21 août 2011 sur le site de Locmariaquer.

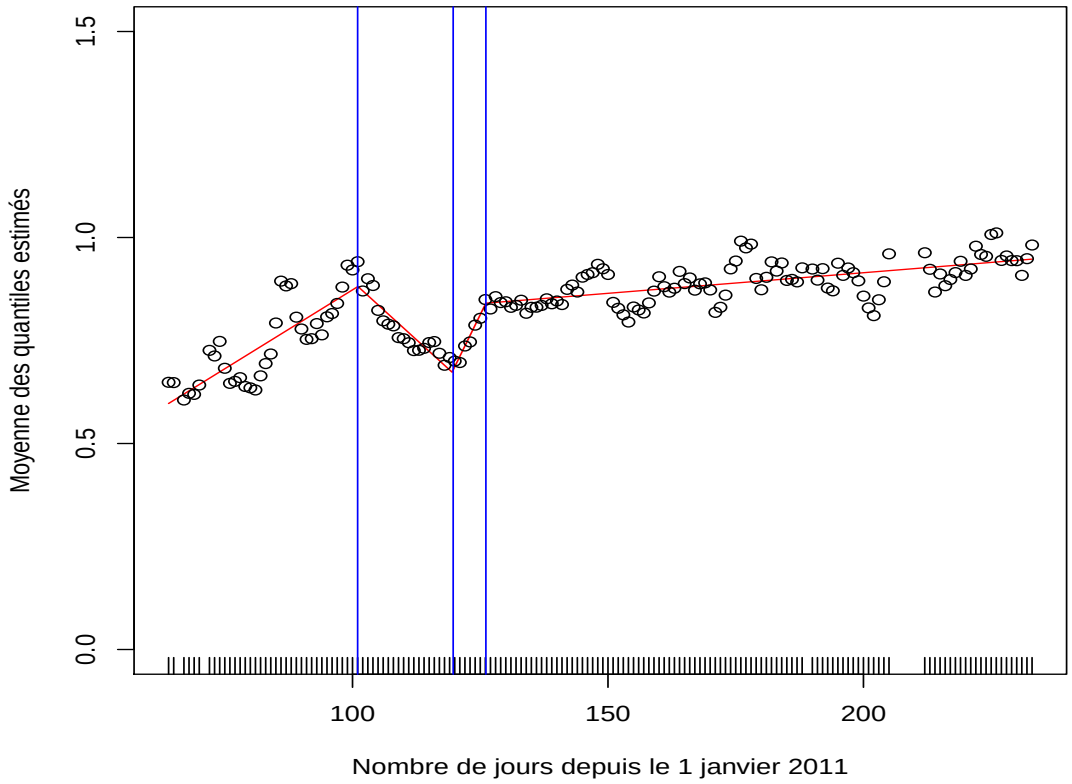

Figure IV.8 – Représentation de l'évolution moyenne des 0.999-quantiles estimés du 3 mars au 21 août sur le site de Locmariaquer. Les droites en rouge correspondent à des régressions par segments avec estimation en brute des points de rupture représentées par les traits verticaux en bleu.

Conclusion générale et perspectives

L'objectif de cette thèse était de proposer des méthodes statistiques basées sur la théorie des valeurs extrêmes pour estimer des probabilités d'évènements rares et des quantiles extrêmes conditionnelles.

Nous donnons la vitesse de convergence des estimateurs du modèle paramétrique ajusté quand le seuil et la taille de la fenêtre sont déterministes. Nous montrons en particulier que la vitesse de convergence obtenue est quasi-optimale dans le cas du modèle de Hall et du modèle de mélange de deux lois de Pareto. Ces résultats de convergence sont également étendus à deux cas : 1) seuil τ adaptif et taille de la fenêtre fixée et 2) taille de la fenêtre adaptative et seuil fixé.

Dans les applications, le seuil τ et la taille de la fenêtre h sont inconnus. Nous proposons donc des procédures du choix de ces deux paramètres. Quand la taille de la fenêtre h est fixée, nous avons proposé une procédure de choix du seuil τ. Quand le seuil τ est fixé, nous avons donné une procédure de choix de la taille de la fenêtre h. En combinant ces deux procédures, nous avons aussi proposé une procédure permettant de déterminer simultanément le seuil τ et la taille de la fenêtre h. En plus, nous avons donné également une procédure de choix de la taille de la fenêtre h par validation croisée.

Nous appliquons la procédure statistique développée sur des données réelles dans un contexte

environnemental. Le but était de déterminer des perturbations environnementales extrêmes en utilisant des mesures à haute fréquance de l'activité des valves des huîtres qui sont considées comme les bioindicateurs de la pollution.

Comme perspectives de ce travail, il serait intéressant d'étudier les propriétées asymptotiques conjointes des estimateurs adaptatifs des deux paramètres τ et h donnés par la procédure du choix simultané du seuil τ et de la fenêtre h. De plus, la consistance faible des estimateurs adaptatifs ou non adaptatifs proposés est établie. Il serait également intéressant d'étudier la normalité asymptotique de ces estimateurs.

Davis et Resnick [65] ont montré que si la distribution est dans le domaine d'attraction de Fréchet ou dans la domaine d'attraction de Gumbel avec un point terminal infini, la queue de la distribution peut alors s'approcher par la queue d'une loi de Pareto (voir Théorème 3.2, [65]). Notre approche pourrait donc s'étendre au cas où la fonction de répartition F_t appartient au domaine d'attraction de Gumbel avec un point terminal infini, en particulier lorsque F_t est une loi à queue de type Weibull.

Dans notre modèle, la covariable T est déterministe. Il serait intéressant d'adapter nos résultats dans le cas où la covariable est aléatoire.

D'un point de vue application, nous prévoyons d'appliquer notre approche sur le traitement de données génomiques du cancer de la prostate dans le cadre d'une collaboration avec l'Université de Toronto au Canada.

Références bibliographiques

[1] F. Marceau. *Contraction of molluscan muscle.* Arch. Zool. Exp. Gen. **2**, 295–469 (1909). 2, 107

[2] K.J.M. Kramer and E.M. Foekema. *The Musselmonitor as biological early warning system : the first decade.* Kluwer Academic Pblishers, New-York (2001). 2, 107

[3] J. Borcherding and M. Volpers. *The dreissena monitor : first results on the application of this biological early warning system in the continuous monitoring of water quality.* Water Science Technology **29**, 199–201 (1994). 2

[4] D. Tran, P. Ciret, A. Ciutat, G Durrieu, and J.-C. Massabuau. *Estimation of potential and limits of bivalve closure response to detect contaminants : application to cadmium.* Environmental Toxicology and Chemistry **22**(4), 914–920 (2003). 4, 33, 34, 35, 110

[5] D. Tran, E. Fournier, G. Durrieu, and J.-C. Massabuau. *Copper detection in the Asiatic clam Corbicula fluminea : Optimum valve closure response.* Aquatic Toxicology **66**, 333–343 (2004). 4, 33, 110

[6] C. Chambon, A. Legeay, G. Durrieu, P. Gonzalez, P. Ciret, and J.-C. Massabuau. *Influence of the parasite worm Polydora sp. on the behaviour of the oyster crassostrea gigas : a study of the respiratory impact and associated oxidative stress.* Marine Biology **152**(2), 329–338 (2007). 4, 34, 110

[7] C. Schwartzmann, G. Durrieu, M. Sow, P. Ciret, C.E. Lazareth, and J.C. Massabuau. *In situ giant clam growth rate behavior : a one-year coupled study of high-frequency noninvasive valvometry and sclerochronology.* Limnology and oceanography **56**(5), 1940–1951 (2011). 4, 110

[8] M. Sow, G. Durrieu, and L. Briollais. *Water quality assessment by means of HFNI valvometry and high-frequency data modeling.* Environmental Monitoring and Assessment **182**(1-4), 155–170 (2011). 4, 33, 35, 37, 110, 115

[9] F. G. Schmitt, M. De Rosa, G. Durrieu, M. Sow, P. Ciret, D. Tran, and J. C. Massabuau. *Statistical analysis of bivalve high frequency microclosing behavior : scaling properties and shot noise modeling.* International Journal of Bifurcation and Chaos **21**(12), 3565–3576 (2011). 4, 33, 110

[10] R. Coudret, G. Durrieu, and J. Saracco. *Comparison of kernel density estimators with assumption on number of modes.* Communication in Statistics - Simulation and Computation **In press** (2013). 4, 33, 35, 37, 110, 115

[11] R. Azaïs, G. Coudret, and G. Durrieu. *A hidden renewal model for monitoring aquatic systems biosensors.* Environmetrics **25**, 189–199 (2014). 4, 33, 110

[12] P. Embrechts, C. Klüppelberg, and T. Mikosh. *Modelling Extremal Events : For Insurance and Finance.* Springer (1997). 7, 15

[13] R. Reiss and M. Thomas. *Statistical Analysis of Extreme Value with applications to assurance, finance, hydrology and other fields.* Springer-Verlag (2007). 7

[14] I. Grama and V. Spokoiny. *Pareto approximation of the tail by local exponential modeling.* Bulletin of Academi of Sciences of Moldova **53**(1), 1–22 (2007). 10, 15

[15] I. Grama and V. Spokoiny. *Statistics of extremes by oracle estimation.* Annals of Statistics **36**(4), 1619–1648 (2008). 10, 15, 24, 43, 50, 54, 57, 58, 83, 84

[16] P. Hall. *On some simple estimates of an exponent of regular variation.* J. Roy. Statist. Soc. B **44**(1), 37–42 (1982). 11, 21

[17] R.L. Smith. *Extreme value analysis of environmental time series : An application to trend detection in ground-level ozone.* Statistical Science **4**(4), 367–393 (1989). 14

[18] A.C. Davison and R.L. Smith. *Models for exceedances over high thresholds.* J. Roy. Statist. Soc. B **52**(3), 393–442 (1990). 14

[19] P. Hall and N Tajvidi. *Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data.* Statistical Science **15**, 153–167 (2000). 14

[20] A. C. Davison and I. Ramesh N. *Local likelihood smoothing of samples extremes.* J. Roy. Statist. Soc. B **62**(1), 191–208 (2000). 14

[21] J. Beirlant and Y. Goegebeur. *Regression with response distributions of pareto-type.* Computational Statistics and Data Analysis **42**, 595–619 (2003). 14

[22] J. Beirlant and Y. Goegebeur. *Local polynomial maximum likelihood estimation for pareto-type distributions.* Journal of Multivariate Analysis **89**, 97–118 (2004). 14, 18

[23] L. Gardes and S. Girard. *A moving window approach for nonparametric estimation of the conditional tail index.* Journal of Multivariate Analysis **99**, 2368–2388 (2008). 14, 15, 23

[24] L. Gardes and S. Girard. *Conditional extremes from heavy -tailed distributuions : an application to the estimation of extreme rainfall return levels.* Extremes **13**, 177–204 (2010). 14

[25] Y. Goegebeur, A. Guillou, and A. Schorgen. *Nonparametric regression estimation of conditional tails : the random covariate case.* Statistics **to appear** (2013). 14

[26] L. Gardes and G. Stupfler. *Estimation of the conditional tail index using a smoothed local hill estimator.* Extremes (to appear). 14

[27] G. Stupfler. *A moment estimator for the conditional extreme-value index.* Electronic Journal of Statictics **7**, 1935–2343 (2013). 14

[28] P. Hall and A. H. Welsh. *Adaptive estimates of parameters of regular variation.* Ann. Statist. **13**(1), 331–341 (1985). 15

[29] H. Drees and E. Kaufmann. *Selecting the optimal sample fraction in univariate extreme value estimation.* Stochastic Process. Appl. **75**, 149–172 (1998). 15

[30] A. Guillou and P. Hall. *A diagnostic for selecting the threshold in extreme-value analysis.* J. Roy. Statist. Soc. Ser. B **63**, 293–305 (2001). 15

[31] R. Huisman, C. G. Koedijk, C. J. M. Kool, and F. Palm. *Tail index estimates in small samples.* Journal of Business and Economic Statistics **19**(2), 208–216 (2001). 15

[32] J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. *Statistics of Extremes : Theory and Applications.* Wiley, Chichester (2004). 15, 17, 63

[33] J. El Methni, L. Gardes, S. Girard, and A. Guillou. *Estimation of extreme quantiles from heavy and light tailed distributions, journal of statistical planning and inference.* Journal of Statistical Planning and Inference **142**(10), 2735–2747 (2012). 15

[34] C.J. Stone. *Optimal global rates of convergence for nonparametric regression.* Ann. Statist. **10**(4), 1040–1053 (1982). 15, 23

[35] M.W. Denny, L.J.H. Hunt, L.P. Miller, and C.D.G. Harley. *On the prediction of extreme ecological events.* Ecological Monographs **79**(3), 397–421 (2009). 16

[36] W. C. Allee, A. E. Emerson, O. Park, T. Park, and K. P. Schmidt. *Principles of animal ecology.* Springer-Verlag, Philadelphia (1949). 16

[37] C. Folke, S. Carpenter, B. Walker, M. Scheffer, T. Elmqvist, L. Gunderson, and C. S. Holling. *Regime shifts, resilience, and biodiversity in ecosystem management.* Annual Review of Ecology and Systematics **35**, 557–581 (2004). 16

[38] J.F. Samain and H. McCombie. *Summer mortality of the Pacific oyster Crassostrea gigas : the Morest Project.* Quae, Versailles (2008). 16

[39] C.A. Burge, L.R. Judah, L.L. Conquest, F.J. Griffin, D. Cheney, A. Suhrbier, B. Vadopalas, P.G. Olin, T. Renault, and C.S. Friedman. *Summer seed mortality of the pacific oyster, crassostrea gigas thunberg grown in tomales bay, california, usa : the influence of oyster stock, planting time, pathogens, and environmental stressors.* Journal of Shelfish Research **26**(1), 163–172 (2007). 16

[40] J. J. Diebolt, Guillou A., and P. Ribereau. *Asymptotic normality of extreme quantile estimators based on the peaks-over-threshold approach.* Comm. Statist. Theory and Methods **36**(5), 869–886 (2007). 18

[41] J. Carreau and S. Girard. *Spatial extreme quantile estimation using a weighted log-likelihood approach.* Journal de la Société Française de Statistique **152**(1), 66–83 (2011). 18

[42] J.G. Staniswalis. *The kernel estimate of a regression function in likelihood-based models.* Journal of the American Statistical Association **84**(405), 276–283 (1989). 18

[43] Clive Loader. *Local regression and likelihood.* Springer (1999). 18

[44] V. Spokoiny. *Multiscale local change point detection with applications to value-at-risk.* Ann. Statist. **37**(3), 1405–1436 (2009). 21, 24, 61, 69

[45] P. Hall and A. H. Welsh. *Best attainable rates of convergence for estimates of parameters of regular variation.* Ann. Statist. **12**(3), 1079–1084 (1984). 21

[46] I. Grama, J.M. Tricot, and J.F. Petiot. *Estimation of the survival probabilities by adjusting a cox model to the tail.* C.R. Acad. Sci. Paris, Ser. I. **349**(13-14), 807–811 (2011). 24

[47] I. Grama, J.M. Tricot, and J.F. Petiot. *Estimation of the extreme survival probabilities from censored data.* Buletinul Academiei De Stiinte a Republicii Moldova. Matematica **74**(1), 33–62 (2014). 24

[48] I. Grama, J.M. Tricot, and J.F. Petiot. *Long term survival probabilities and Kaplan-Meier estimator.* Proceedings of the Joint Statistical Meeting, Montreal, Canada, 2013 (2013). 24

[49] D. Tran, E. Fournier, G. Durrieu, and J.-C. Massabuau. *Inorganic mercury detection by valve closure response in the freshwater clam corbicula fluminea : integration of time and water metal concentration changes.* Environmental Toxicology and Chemistry **26**, 1545–1551 (2007). 33, 110

[50] F. G. Doherty, D. S. Cherry, and J. Cairns. *Valve closure responses of the asiatic clam c. fluminea exposed to cadmium and zinc.* Hydrobiologia **153**, 159–167 (1987). 33

[51] K. Nagai, T. Honjo, J. Go, H. Yamashita, and S.J. Oh. *Detecting the shellfish killer heterocapsa circularisquama (dinophyceae) by measuring the bivalve valve activity with hall element sensor.* Aquaculture **255**, 395–401 (2006). 33

[52] A. Robson, R. Wilson, and C. Garcia de Leaniz. *Mussels flexing their muscles : a new method for quantifying bivalve behavior.* Marine Biology **151**, 1195–1204 (2007). 33

[53] L.J. Jou and C.M. Liao. *A dynamic artificial clam (corbicula fluminea) allows parcimony on-line measurement of waterborne metals.* Environmental Pollution **144**, 172–183 (2006). 33

[54] R Development Core Team. *R : A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, (2012). ISBN 3-900051-07-0. 35

[55] D. Tran, H. Haberkorn, P. Soudant, P. Ciret, and J.C. Massabuau. *Behavioral responses of Crassostrea gigas exposed to the harmful algae Alexandrium minutum.* Aquaculture **298**, 338–345 (2010). 35

[56] G. Durrieu, I. Grama, Q.K. Pham, and J.M. Tricot. *Nonparametric adaptive estimation of conditional probabilities of rares events and extreme quantiles.* Extremes **Submitted** (2014). 68, 84

[57] O. V. Lepski. *One problem of adaptive estimation gaussian white noise.* Theory of Probab. Appl. **35**, 459–470 (1990). 72

[58] L. Gardes and S. Girard. *Conditional extremes from heavy tailed distributions : an application to the estimation of extreme rainfall return levels.* Extremes **13**, 177–204 (2010). 101

[59] D. Fawcett and D. Walshaw. *Improved estimation for temporally clustered extremes.* Environmetrics **18**(2), 173–188 (2007). 101

[60] J. Borcherding. *Ten years of practical experience with the dreissena-monitor, a biological early warning system for continuous water quality monitoring.* Hydrobiologia **556**, 417–426 (2006). 107

[61] E. Fournier, D. Tran, F. Denison, J.-C. Massabuau, and J. Garnier-Laplace. *Valve closure response to uranium exposure for a freshwater bivalve corbicula fluminea : quantification of the influence of ph.* Envir. Tox. Chem **23**, 1108–111 (2004). 110

[62] D. Tran, A. Nadau, G. Durrieu, P. Ciret, J.-P. Parisot, and J.-C. Massabuau. *Field chronobiology of a molluscan bivalve : How the moon and the sun cycles interact to drive oyster activity rhythms.* Chronobiology International **28**(4), 307–317 (2011). 110

[63] C.M. Liao, L.J. Jou, and B.C. Chen. *Risk-based approach to appraise valve closure in the clam corbicula fluminea in response to waterborne metals.* Environ. Pollut. **135**, 41–52 (2005). 110

[64] C.M. Liao, C.M. Lin, L.J. Jou, and K.C. Chiang. *Linking valve closure behavior and sodium transport mechanism in freshwater clam corbicula fluminea in response to copper.* Environ. Pollut. **147**, 656–667 (2006). 110

[65] R. Davis and S. Resnick. *Tail estimates motivated by extreme value theory.* The Annals of Statistics **42**, 1467–1487 (1984). 120